IA^2 Institut d'Automne en Intelligence Artificielle - Le Touquet

IA responsable : transparence, explicabilité et impacts sociétaux

Céline Hudelot

23 octobre 2026



CentraleSupélec | université PARIS-SACLAY

Bio



Prof. Céline Hudelot, Informatique
Directrice du laboratoire MICS
Recherche en IA pour l'analyse sémantique de données non-structurées - Apprentissage - Représentation de la connaissance et du raisonnement. https://scholar.google.fr/citations?user=gFlAh6MAAAAJ&hl=fr

1. Intelligences artificielles
2. IA responsable
3. Explicabilité - Interprétabilité
   - Attributions
   - Attribution : définition et vue unifiée
   - Explicabilité par concepts
4. Transparence
5. Conclusion

Intelligence artificielle : définition d'un père fondateur

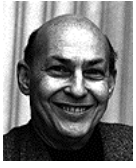
1956 : le terme ≪ **intelligence artificielle** ≫ naît lors de la conférence de Dartmouth.



Marvin Minsky:
"Science qui consiste à faire faire aux machines ce que l'homme ferait moyennant une certaine intelligence"

Une définition relative **à ce que savent faire les hommes**

- Une certaine intelligence?
- Quelle est la portée de ≪ faire faire ≫ ?



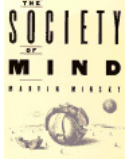
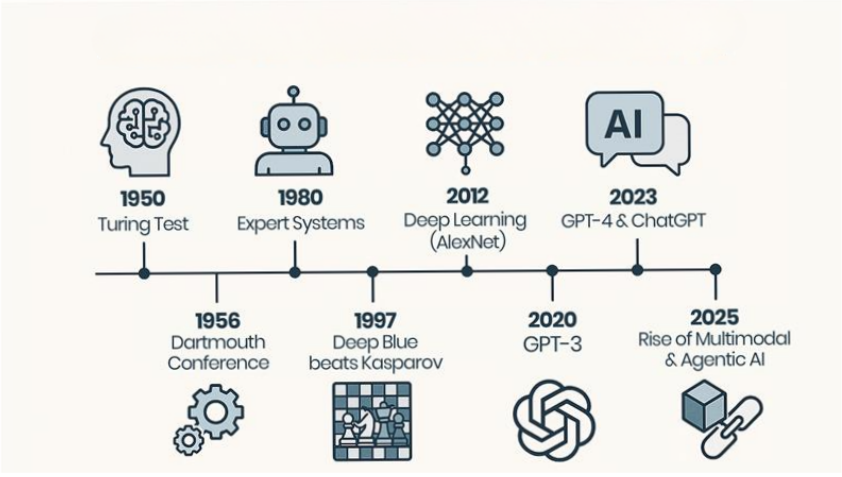


Figure: Minsky : Interaction entre de nombreux agents

Mais très vite, paradoxe de Moravec.

Intelligence artificielle : 70 ans d'histoire et d'avancées

Avec plusieurs changements de paradigme majeurs ces dernières années

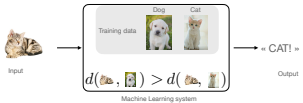


Source : https://www.linkedin.com/pulse/history-artificial-intelligence-timeline-innovation-from-bansal-cpa-asw6f/

Intelligences artificielles : deux approches principales (et antagonistes[1])

Deux hypothèses différentes

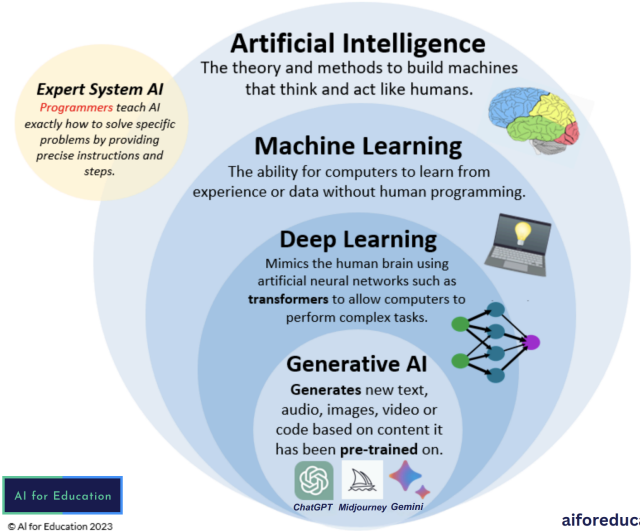
- **Connaissances et raisonnement** sont complexes mais implicitement dans les données.
  - **IA centrée sur les données** - Apprentissage à partir de données - Approches connexionnistes.
  - Exploitation de l'expérience représentée par des données (annotées), pour construire des modèles prédictifs calibrés à partir de celles-ci.
- **Connaissance et raisonnement humain** peuvent être capturés, même si partiellement incomplets : représentation explicite de la connaissance.
  - **IA symbolique** - Basée sur la modélisation formelle du raisonnement, sur les formalismes de représentation de la connaissance et du raisonnement.





[1] D. Cardon et al - La Revanche des neurones - https://hal.archives-ouvertes.fr/hal-02005537/document

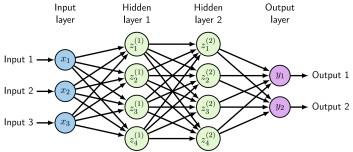
Intelligences artificielles : plusieurs approches



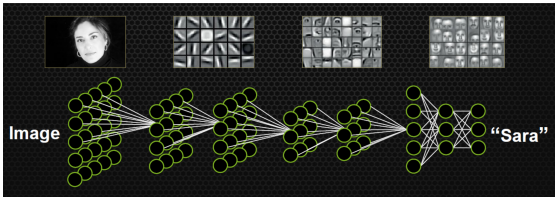
2010 : l'avènement de l'apprentissage profond

Réseau de Neurones Profonds

$\Longrightarrow Y = \psi(X)$, avec $\psi(X) = h_M \circ g_M \circ \ldots \circ h_1 \circ g_1(X)$ où h_i des transformations non-linéaires et g_i des transformations affines.



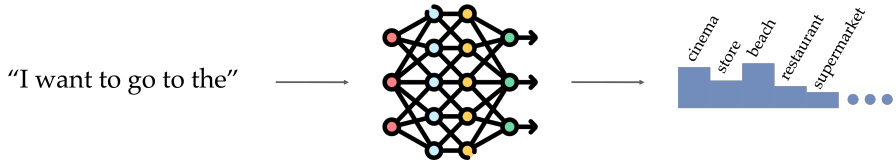
Apprentissage de représentations et transfert d'apprentissage



Chaque couche transforme les données brutes en **représentations hiérarchiques**. Apprentissage d'un **espace latent** où la tâche devient plus simple [Bengio et al,2014]. Les représentations apprises sont **transférables**.

2020 : IA Générative et LLMs

Un LLM est un réseau neuronal conçu pour traiter et générer du texte humain.

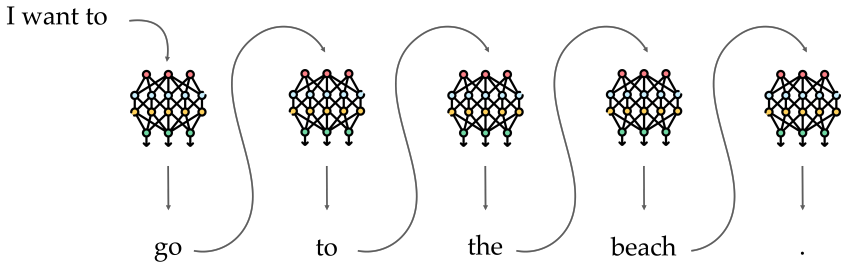


- Le modèle (un réseau de neurones) est entraîné à grande échelle avec comme tâches MLM (Masked Language Modelling), i.e. prédire des tokens cachés dans son entrée ou CLM (Causal Language Modelling), i.e. prédire le token suivant
- Cet apprentissage se fait de manière auto-supervisée (self-supervised learning) sur un très gros corpus (e.g. le WEB)
- Cette phase permet au modèle d'acquérir de nombreuses *connaissances* et capacités.
- À l'inférence (utilisation du modèle), le LLM prédit, étant donnée une entrée appelée prompt, une distribution de probabilité pour les mots suivants étant donné un vocabulaire de mots ou de tokens.

Slide credit : P. Colombo

2020 : IA Générative et LLMs

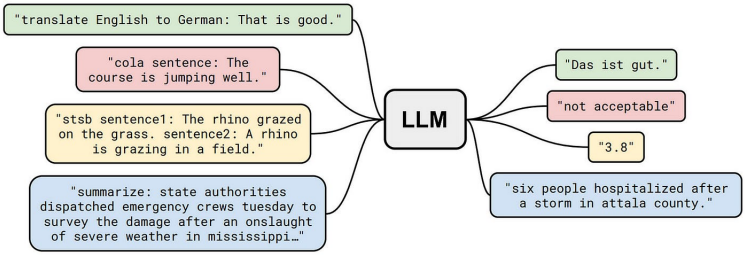
Comment un LLM génère du texte ? Il suit une approche progressive pour générer du texte, en commençant par un préfixe, puis en générant un mot à la fois.



Slide credit : P. Colombo

2020 : IA Générative et LLMs

Comment tirer parti des LLM pour n'importe quelle tâche ? **Prompting**.

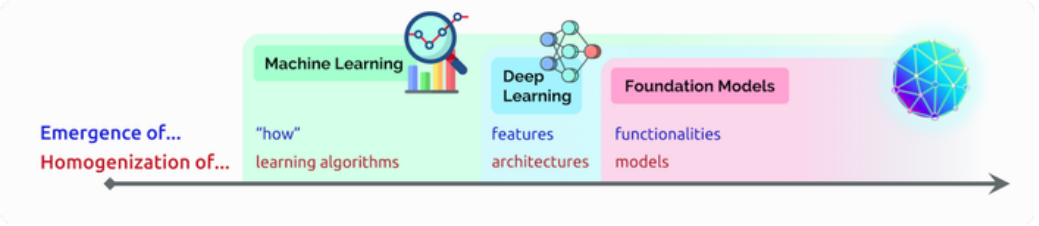


Source : [Raffel et al, JMLR 2020] Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

2020 : le paradigme des modèles de fondation

Un nouveau paradigme pour la construction des systèmes d'IA.

- Modèle de fondation : modèle de grande taille, entraîné sur un grand volume de données non étiquetées, adaptable à un large éventail de tâches.
- Des modèles d'IA à usage général (AGI : Artificial General Intelligence, GPAI : General-Purpose AI)



Source :[Bommasani et al, 2022] On the Opportunities and Risks of Foundation Models

Quelques modèles de fondation

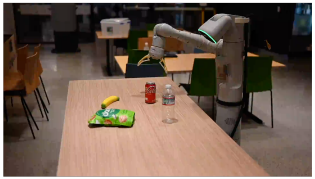
Segment Anything by Meta AI



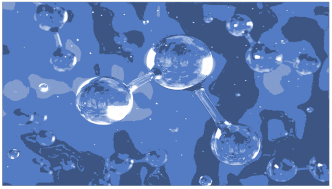




RT-2: Vision-Language-Action Models by Google Deep Mind

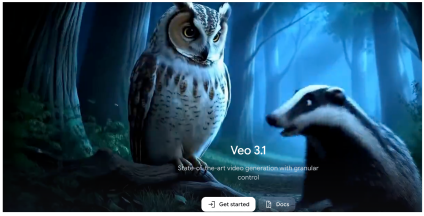


MoLFormers, foundation models on chemicals by IBM



- https://chatgpt.com/
- https://segment-anything.metademolab.com/
- https://robotics-transformer2.github.io/
- https://molformer.res.ibm.com/

Quelques modèles de fondation



Video generation : Veo 3.x, Kling

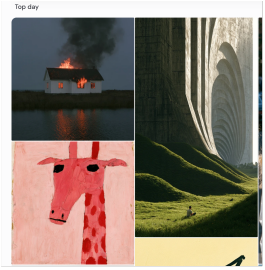
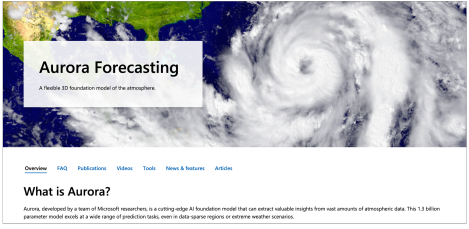
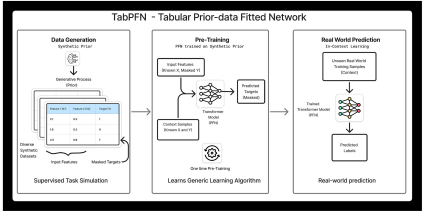


Image generation : Mid journey,
Nano Banana Pro, FLUX (modèleeuropéen)

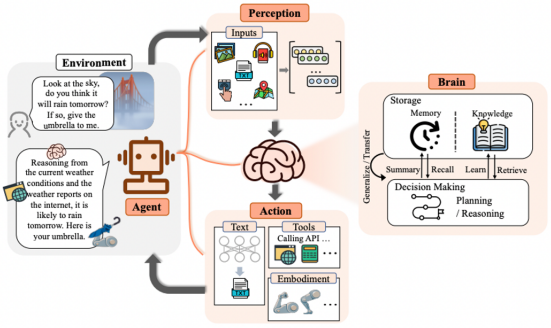


Modèle de fondation de l'atmosphère



Données tabulaires

Plus récemment, l'IA agentique



Un agent est un système qui exploite un modèle d'IA (souvent un LLM) pour interagir avec son environnement afin d'atteindre un objectif défini par l'utilisateur. Il combine le **raisonnement**, la **planification** et l'**exécution d'actions** (souvent via des outils externes) pour accomplir des tâches, i.e. programme dans lequel les résultats générés par les LLMs contrôlent le déroulement des tâches.

Agents à base de LLMs

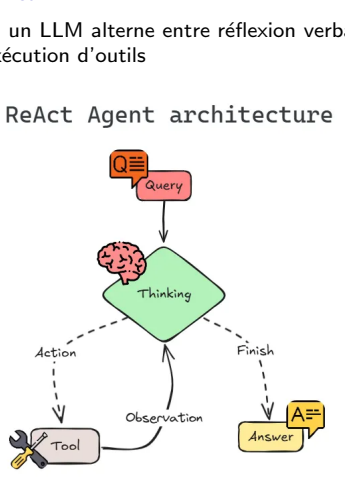
LMM placé dans une boucle **perception → raisonnement → action → observation.**
On ajoute au modèle :

- **des outils** : appels d'API, exécution de code, recherche web, manipulation de fichiers, ou contrôle d'une interface graphique;
- **une mémoire** : le contexte de travail, et éventuellement une mémoire longue externe ;
- **de la planification** : décomposition de la tâche, réflexion sur les erreurs, replanification ;
- **un critère d'arrêt** : la tâche est accomplie, un budget est épuisé, ou il faut demander à l'humain.

Agents à base de LLMs

Le pattern ReAct

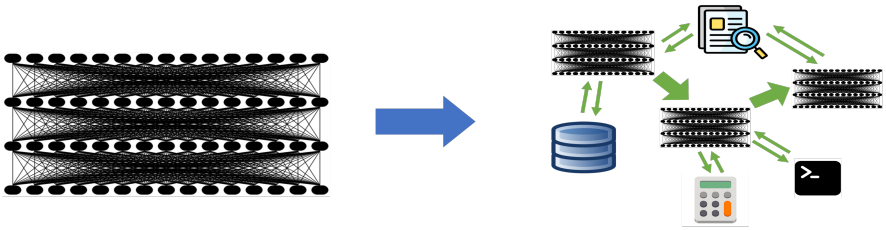
Paradigme où un LLM alterne entre réflexion verbale explicite et exécution d'outils



- **Thought** : Le LLM évalue la situation, décompose le problème et décide de la prochaine étape
- **Action** : L'agent sélectionne un outil externe et formule un appel d'API ou une requête spécifique.
- **Observation** :L'agent lit le résultat retourné par l'outil pour enrichir sa mémoire avant le tour suivant.

[Yao et al., 2023] ReAct: Synergizing Reasoning and Acting in Language Models

Systèmes multi-agents

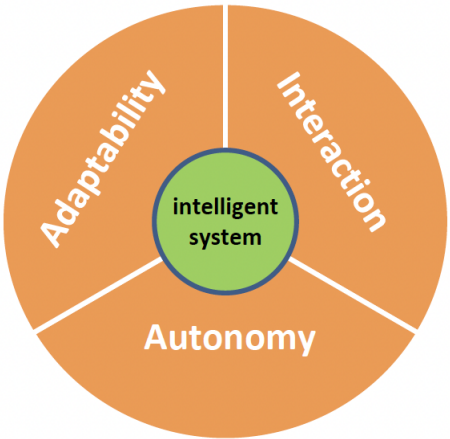


- Pas un seul système monolithique mais plusieurs composants.
- **Bénéfices** : amélioration grâce à la conception de systèmes, à l'aspect dynamique et à un contrôle facilité.
- Mais la conception, l'optimisation et l'opérationnalisation doivent être menées avec soin.

L'IA aujourd'hui

Systèmes intelligents dans le monde réel

Des systèmes en **interaction** avec le monde, **autonomes** et en passe de **s'adapter**.



L'IA aujourd'hui ... partout !



L'IA aujourd'hui ... très impressionnante



Robot Taxi - Waymo







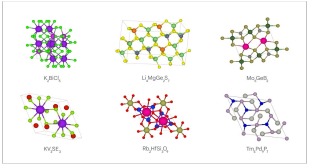

V1 (Feb 22) V2 (Apr 22) V3 (Jul 22) V4 (Nov 22) V5 (Mar 23) V6 (Dec 23)



Génération de vidéos



Génération de codes

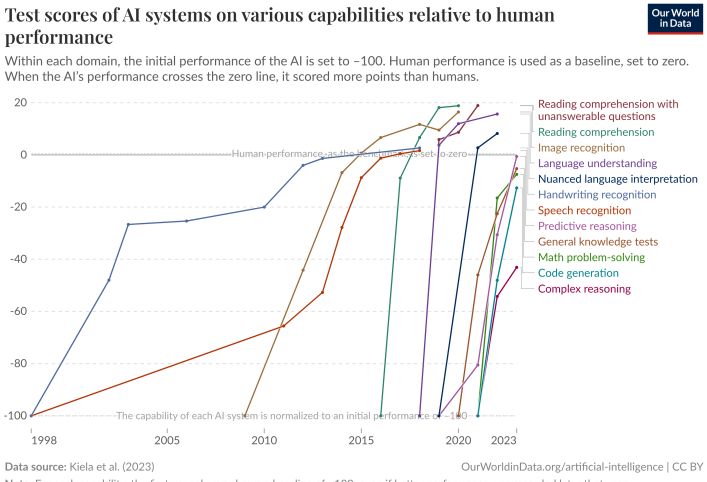




Découverte de nouveaux matériaux

L'IA aujourd'hui ... très impressionnante

Vers l'AGI ?

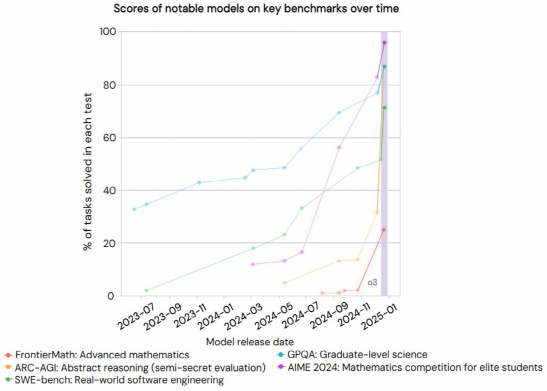


À toute vitesse et avec accélération

Major advances at a very fast pace !

International AI Report [Bengio al, 25]

« Between the end of the writing period this report (5 December 2024) and the publication of this report in January 2025 an important development took place »



Capable de créer de l'art !



Figure: The next Rembrandt: `https://www.nextrembrandt.com/`

Ce qui soulève des questions de propriété intellectuelle : qui est l'auteur ?

Capable de faire avancer les connaissances !



Nouveau coup de massue d'OpenAI sur les mathématiciens : « Ils ont annihilé mon domaine de recherche »

Après les dernières annonces de l'entreprise américaine sur la résolution de plusieurs centaines de problèmes de maths, le médaillé Fields Hugo Duminil-Copin évoque un « démantèlement » de sa discipline. Il craint en particulier les conséquences de l'intelligence artificielle sur l'avenir des jeunes.

Par David Larousserie

Publié le 10 octobre 2026 à 09h30, modifié à 17h45 · Lecture 3 min. · Read in English

Article réservé aux abonnés

Offrir l'article · Lire plus tard

Le mathématicien français Hugo Duminil-Copin, à Genève, le 28 juin 2023. FABRICE COFFRINI/AFP

« Mardi, dans la nuit, j'ai pleuré et cela a réveillé ma fille », a confessé le mathématicien Hugo Duminil-Copin, médaille Fields 2022, lors d'une visioconférence organisée par la Société mathématique de France, jeudi 8 octobre. La raison de ses larmes ? L'entreprise américaine d'intelligence artificielle (IA) OpenAI a mis en ligne plus de 719 manuscrits, répartis en 372 familles, prétendant résoudre de nombreux problèmes mathématiques encore ouverts. Pas moins de 16 questions qui intéressaient le chercheur français sont concernées.

Édition du jour

Date du mardi 13 octobre

Le Monde

Accord Trump-Poutine : l'Ukraine et l'Europe délitées

Lire le journal numérique

Lire les éditions précédentes

PUBLICITÉ

nature

Explore content · About the journal · Publish with us · Subscribe

nature > news > article

NEWS · 18 September 2026

AI cracked the Navier–Stokes challenge. What does that mean for physics?

Physicists and mathematicians are going beyond the classic equations of fluid dynamics to understand turbulence — often with the help of AI.

By Nicola Jones

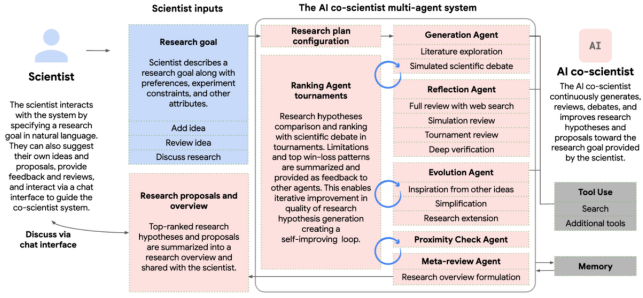
Save article

Chaotic vortices the wake of an aeroplane can contribute to aerodynamic drag. Credit: Photo123/JGI/Getty

Last week, OpenAI stunned the world by claiming that its most advanced artificial-intelligence model had solved a Millennium Prize Problem: one of just a handful of notoriously tricky mathematical conundrums that each come with a US$1-million prize. The AI solution proved mathematically that there are instances in which the Navier–Stokes equations — famous, nineteenth-century formulae that model how fluids flow — produce a 'singularity', predicting physically impossible infinite speeds for gases or liquids.

En 2024, des IAs médailles d'argent aux OIM, en 2025, médailles d'or et en 2026, score parfait (tous les problèmes)

Des agents scientifiques



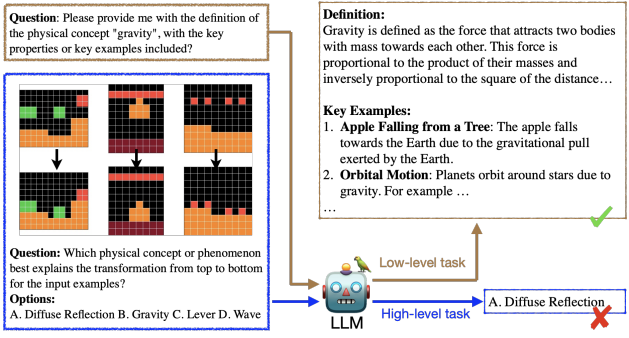
The AI co-scientist system design

Figure:
https://research.google/blog/accelerating-scientific-breakthroughs-with-an-ai-co-scientist/

Mais,

Les systèmes d'IA sont-ils plus que des « perroquets stochastiques » ?

[Bender et al.] [2] Comprennent-ils vraiment les concepts qu'ils utilisent ?



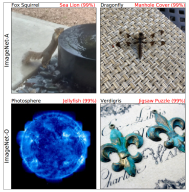
The Stochastic Parrot on LLM's Shoulder: A Summative Assessment of Physical Concept Understanding
`https://physico-benchmark.github.io/`

[2] `https://dl.acm.org/doi/pdf/10.1145/3442188.3445922`

Sont-ils robustes et sûrs ?



[Goodfellow el al,] Explaining and Harnessing adversarial examples



Attaques adverariales naturelles (Hendrycks et al, 2021) Natural Adversarial Examples

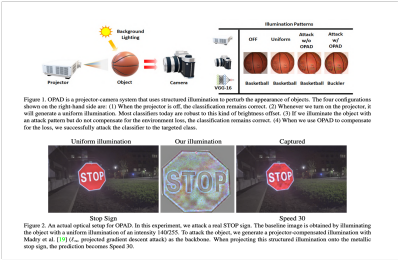


Figure 1. OPAD is a projector-camera system that uses structured illumination to perturb the appearance of objects. The four configurations shown on the right-hand side are: (1) When the projector is off, the classification remains correct. (2) Whenever we turn on the projector, it will generate a uniform illumination. Most classifiers today are robust to this kind of brightness offset. (3) If we illuminate the object with an attack pattern but do not compensate for the environment loss, the classification remains correct. (4) When we use OPAD to compensate for the loss, we successfully attack the classifier to the targeted class.

Figure 2. An actual optical setup for OPAD. In this experiment, we attack a real STOP sign. The baseline image is obtained by illuminating the object with a uniform illumination of an intensity 140/255. To attack the object, we generate a projector-compensated illumination with Madry et al. [19] (ℓ_∞, projected gradient descent attack) as the backbone. When projecting this structured illumination onto the metallic stop sign, the prediction becomes Speed 30.

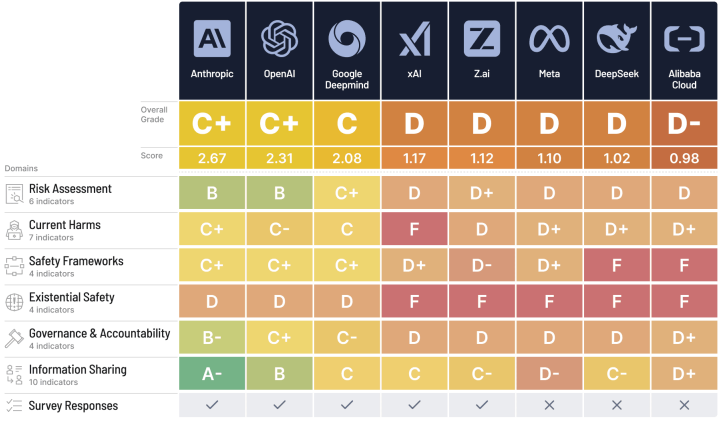
Des attaques physiques [Gnanasambandam et al, 2021 – Optical Adversarial Attack

MITRE ATLAS, Adversarial Threat Landscape

Base de connaissances sur les tactiques et techniques adverses, et études de cas pour les systèmes d'apprentissage automatique (ML) : `https://atlas.mitre.org/`

Sont-ils sûrs ?

AI Safety Index - FLI institute. L'indice de sécurité de l'IA du Future of Life Institute propose une évaluation indépendante des efforts déployés par sept grandes entreprises du secteur de l'IA pour gérer à la fois les dangers immédiats et les risques catastrophiques liés aux systèmes d'IA avancés.



	Anthropic	OpenAI	Google Deepmind	xAI	Z.ai	Meta	DeepSeek	Alibaba Cloud
Overall Grade	C+	C+	C	D	D	D	D	D-
Score	2.67	2.31	2.08	1.17	1.12	1.10	1.02	0.98
Domains								
Risk Assessment (6 indicators)	B	B	C+	D	D+	D	D	D
Current Harms (7 indicators)	C+	C-	C	F	D	D+	D+	D+
Safety Frameworks (4 indicators)	C+	C+	C+	D+	D-	D+	F	F
Existential Safety (4 indicators)	D	D	D	F	F	F	F	F
Governance & Accountability (4 indicators)	B-	C+	C-	D	D	D	D	D+
Information Sharing (10 indicators)	A-	B	C	C	C-	D-	C-	D+
Survey Responses	✓	✓	✓	✓	✓	×	×	×

Future of Life Institute | AI Safety Index: Winter 2025

Grading: We're using the US GPA system for grade boundaries: A+, A, A-, B+, [...], F letter values correspond to numerical values 4.3, 4.0, 3.7, 3.3, [...], 0.

Quels sont les risques ?

Décomposition des risques en 3 familles

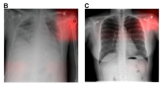
- **Risques d'usage malveillant** : un individu ou un groupe utilise intentionnellement l'IA à des fins nuisibles.
- **Risques d'alignement** : liés aux systèmes d'IA eux-mêmes, en raison de problèmes de conception inhérents, tels que des systèmes poursuivant des objectifs non alignés sur les valeurs humaines.
  - **Spécification** : nous n'avons pas correctement spécifié ce que nous voulions.
  - **Généralisation** : même avec une bonne spécification, le système d'IA peut apprendre et poursuivre quelque chose de différent de ce qui était prévu.
  - **Sous-objectifs** : en poursuivant ses objectifs appris, le système peut développer des sous-objectifs problématiques, tels que de s'empêcher d'être éteint.
- **Risques systémiques** : problèmes plus larges qui émergent lorsque l'on considère l'IA non pas isolément, mais comme une variable parmi d'autres dans l'interaction globale des incitations au sein de systèmes complexes tels que la politique, la société ou l'économie, où aucune entité unique n'est responsable.

Mais aussi différents niveaux de gravité.

Sont-ils dignes de confiance ?

CNNs learn to predict pneumonia by detecting hospital which took the image

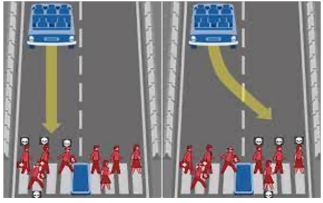
- Study on detecting pneumonia using 158,323 chest radiographs
- CNNs robustly identified hospital system and department within a hospital
- CNN has learned to detect a metal token that radiology technicians place on the patient in the corner of the image

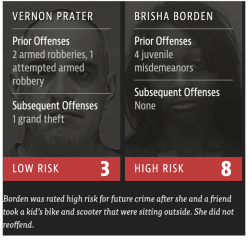


Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study.
Zech JR1, Badgeley MA2, Liu M2, Costa AB3, Titano JJ4, Oermann EK3. https://www.ncbi.nlm.nih.gov/pubmed/30399157

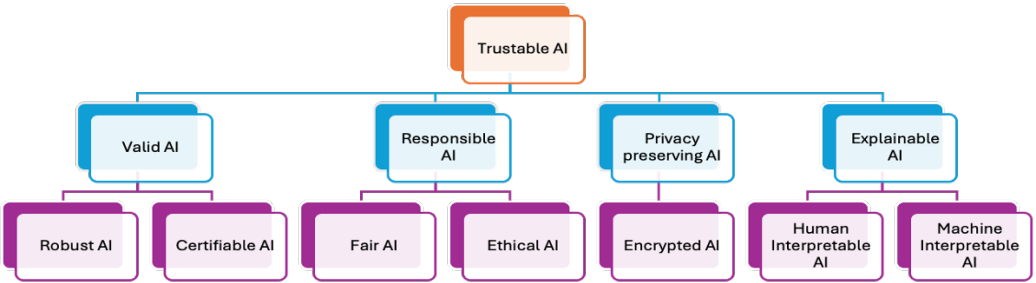








Manifestes pour des IAs dignes de confiance



IA Digne de Confiance

IA Digne de Confiance (*Trustworthy AI*)

Une IA digne de confiance comporte les 3 éléments suivants, devant être présents tout au long du cycle de vie du système.

1. **licite** - en assurant le respect des législations et réglementations applicables.
2. **éthique** - en assurant l'adhésion à des valeurs éthiques.
3. **robuste** - sur le plan technique et social pour éviter de causer des préjudices involontaires.

Source : *European Commission's High-Level Expert Group on AI*

Pour aller plus loin - Rapport *Lignes directrices en matière d'éthique pour une IA Digne de Confiance*[3]

[3] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

IA Digne de Confiance

Trois aspects

1. Les Fondements d'une IA digne de confiance : définition de **principes**
2. La réalisation d'une IA digne de confiance : définition et mise en œuvre d'**exigences essentielles** techniques ou non-techniques.
3. Évaluation d'une IA digne de confiance : concrétisation d'une IA digne de confiance, **liste d'évaluation** alignée sur la structure de gouvernance.

IA Digne de Confiance : les fondements

Se basent sur les droits fondamentaux : (1) Respect de la dignité humaine; (2) Liberté des individus; (3) Respect de la démocratie, de la justice et de l'État de droit; (4) Égalité, non-discrimination et solidarité; et (5) Droit des citoyens.

4 principes (impératifs) éthiques pour l'IA

1. **Respect de l'autonomie humaine**
   Les êtres humains qui interagissent avec des systèmes d'IA doivent être en mesure de conserver leur autodétermination totale et effective et de participer au processus démocratique.
2. **Prévention de toute atteinte**
   Les systèmes d'IA ne devraient ni porter atteinte, ni aggraver une atteinte, ni nuire aux êtres humains de quelque manière que ce soit.
3. **Équité**
   La mise au point, le déploiement et l'utilisation de systèmes d'IA doivent être équitables.
4. **Explicabilité**
   Renforcer et conserver la confiance des utilisateurs envers les systèmes d'IA

IA Digne de Confiance : les exigences

Traduction des principes en exigences concrètes s'appliquant aux différentes parties prenantes : concepteurs, prestataires et utilisateurs finaux, ainsi que la société au sens large

7 exigences

1. **Action humaine et contrôle humain** : *soutenir l'autonomie et la prise de décision humaine, permettre le contrôle humain.*
2. **Robustesse technique et sécurité** : *Résilience aux attaques et sécurité, plans de secours et sécurité générale, précision, fiabilité et reproductibilité*
3. **Respect de la vie privée et gouvernance des données** : *protection, qualité, intégrité et accès aux données*
4. **Transparence** : *explicabilité, tracabilité, communication (l'anti-test de Turing).*
5. **Diversité, non-discrimination et équité** : *absence de biais injustes, systèmes centrés sur l'utilisateur, accessibles*
6. **Bien-être sociétal et environnemental** : *durabilité et responsabilité écologique des systèmes d'IA, sans incidences sociales néfastes*
7. **Responsabilité** : *auditabilité, réduction et documentation des incidences négatives, arbitrages, recours*

Un sujet d'enjeu!

- **Pour les grands acteurs de l'IA et de la donnée.**
  - e.g Google Responsible AI principles and practices[4]
  - e.g. Microsoft AI principles[5]
- **Pour les institutions (au sens large)**
  - e.g. Union Européenne : Lignes directrice en matière d'éthique pour une IA Digne de Confiance.
  - e.g. des instituts dédiés
    - The Institute for Ethical AI et Machine Learning[6]
    - Programme Operational AI Ethics, Télécom ParisTech[7]
    - Grand défi du Conseil de l'Innovation : Confiance.AI : Sécuriser, certifier et fiabiliser les systèmes fondés sur l'intelligence artificielle[8]
- **Pour la Société**
  - e.g. Unesco : Elaboration of a Recommendation on the ethics of artificial intelligence[9]

[4] https://ai.google/responsibilities

[5] https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1:primaryr6

[6] https://ethical.institute/

[7] https://www.telecom-paris.fr/fr/recherche/axes-strategiques/sciences-des-donnees-et-intelligence-artificielle/operational-ai-ethics

[8] https://www.gouvernement.fr/grand-defi-securiser-certifier-et-fiabiliser-les-systemes-fondes-sur-l-intelligence-artificielle

[9] https://en.unesco.org/artificial-intelligence/ethics

AI act

Construit sur les risques.

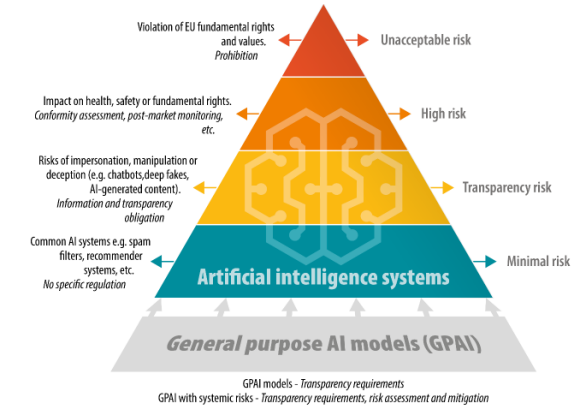


Figure: Source :ethical.institute

IA Responsable

IA responsable

Definition - Stanford HAI

Ensemble des cadres, principes et pratiques qui guident le développement et le déploiement éthiques et sûrs des systèmes d'intelligence artificielle. Concevoir les technologies d'IA en tenant compte de l'**équité**, de la **transparence**, de la **responsabilité**, de la **vie privée** et de l'**impact sociétal potentiel**.

Definition [Dignum, 2019][a]

[a] V. Dignum - Responsible Artificial In telligence

L'IA responsable concerne la **responsabilité humaine** dans le développement des systèmes intelligents, selon des principes et des valeurs humaines fondamentales, afin d'assurer l'épanouissement et le bien-être humains dans un monde durable.

IA Responsable [Dignum, 2019]

Les grands acteurs

- L'IA comme système socio-technique.
  - **La technique** : Données, modèles, algorithmes, infrastructures.
  - **Les personnes** : Concepteurs, décideurs, utilisateurs, personnes affectées
  - **Les institutions** : Organisations, normes, droit, gouvernance.

IA Responsable [Dignum, 2019]

Trois niveaux

- **Ethics in design**
  - Méthodes pour intégrer les valeurs dans le processus de conception
- **Ethics by design**
  - Raisonnement éthique intégré au comportement du système
  - Le cadre ART : Accountability · Responsibility · Transparency
- **Ethics for design(ers)**
  - Intégrité et codes de conduite des praticiens

Le cadre ART : Accountability · Responsibility · Transparency - [Dignum, 2019]

A

Accountability

Pouvoir expliquer et justifier les décisions du système.

R

Responsibility

Le rôle des humains et la chaîne de responsabilité.

T

Transparency

Pouvoir inspecter mécanismes, données et décisions.

Le cadre ART : Accountability · Responsibility · Transparency - [Dignum, 2019]

Accountability : challenges

- Expliquer et justifier les décisions auprès des personnes concernées
  - **"Optimal AI is explainable AI"**
  - Rattacher chaque décision aux valeurs et normes qui la fondent
  - Rendre compte du processus de conception, pas seulement du résultat

Le cadre ART : Accountability · Responsibility · Transparency - [Dignum, 2019]

Responsability : challenges

- La responsabilité incombe aux personnes et aux organisations, jamais à la machine. Le système est un outil, pas un agent moral et donc **une chaine de responsabilité**



- Niveaux d'autonomie :
  - Opérationnelle : actions /plans
  - Décisionelle : Buts / Motivations
  - Autonomie réalisable : en fonction du contexte et de la complexité de la tâche

Le cadre ART : Accountability · Responsibility · Transparency - [Dignum, 2019]

Responsability : challenges

Pouvoir inspecter, pas seulement expliquer



Données

Provenance, collecte, biais connus



Mécanismes

Modèles, algorithmes, limites



Décisions

Ce qui a été produit et pourquoi



Gouvernance

Qui a choisi quoi, selon quels critères

La transparence est une condition de l'accountability : on ne peut justifier que ce qui peut être examiné.

Le cadre ART : Accountability · Responsibility · Transparency - [Dignum, 2019]

Des principes aux outils

Accountability

- Journalisation des décisions
- Audits internes et externes
- Voies de recours pour les personnes affectées

Responsibility

- Rôles explicites (qui décide, qui valide)
- Supervision humaine
- Revue éthique des projets

Transparency

- Datasheets for datasets
- Model cards
- Documentation des choix de conception

Écho européen : transparence et accountability figurent parmi les sept exigences des lignes directrices du HLEG pour une IA digne de confiance (2019).

Explicabilité - Interprétabilité

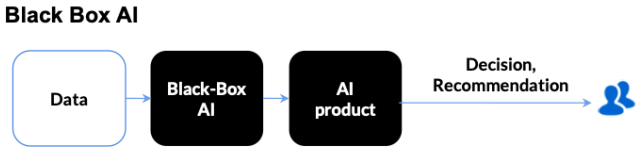
eXplainable Artificial Intelligence (XAI)

Définition

L'explicabilité en IA vise à mettre au point un ensemble de techniques pour concevoir des modèles plus explicables, tout en conservant un niveau de performance élevé. Il s'agit de **rendre leur comportement plus intelligible pour les humains** en leur fournissant des explications. Un système XAI doit être capable:

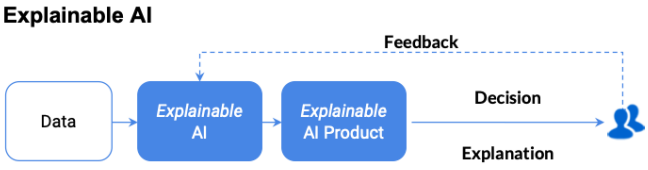
- d'expliquer ses capacités et sa compréhension de l'environnement,
- d'expliquer ce qu'il a fait, ce qu'il fait, ce qu'il fera ensuite,
- de mettre en évidence les informations pertinentes qu'il utilise.

eXplainable Artificial Intelligence (XAI)



Confusion with Today's AI Black Box

- Why did you do that?
- Why did you not do that?
- When do you succeed or fail?
- How do I correct an error?



Clear & Transparent Predictions

- I understand why
- I understand why not
- I know why you succeed or fail
- I understand, so I trust you

Motivations

Plusieurs réponses

- Renforcer la confiance dans les prédictions du modèle.



- Bryce Goodman et al. European union regulations on algorithmic decision-making and a"right to explanation".
- Finale Doshi-Velez et al. Accountability of ai under the law: The role of explanation
- Gabriel Cadamuro et al. Debugging machine learning models
- Alfredo Vellido et al. Making machine learning models interpretable
- [Ribeiro et al,. 2016] Why Should I Trust You?" Explaining the Predictions of Any Classifier
- [Labarta et al, 26] From Attribution to Action: A Human-Centered Application of Activation Steering

Motivations

Plusieurs réponses

- Renforcer la confiance dans les prédictions du modèle.
- Aider à satisfaire les exigences réglementaires et le processus de certification (e.g. AI Act)



- Bryce Goodman et al. European union regulations on algorithmic decision-making and a"right to explanation".
- Finale Doshi-Velez et al. Accountability of ai under the law: The role of explanation
- Gabriel Cadamuro et al. Debugging machine learning models
- Alfredo Vellido et al. Making machine learning models interpretable
- [Ribeiro et al,. 2016] Why Should I Trust You?" Explaining the Predictions of Any Classifier
- [Labarta et al, 26] From Attribution to Action: A Human-Centered Application of Activation Steering

Motivations

Plusieurs réponses

- Renforcer la confiance dans les prédictions du modèle.
- Aider à satisfaire les exigences réglementaires et le processus de certification (e.g. AI Act)
- Révéler les biais ou d'autres effets involontaires d'un modèle.



(a) Husky classified as wolf



(b) Explanation

- Bryce Goodman et al. European union regulations on algorithmic decision-making and a"right to explanation".
- Finale Doshi-Velez et al. Accountability of ai under the law: The role of explanation
- Gabriel Cadamuro et al. Debugging machine learning models
- Alfredo Vellido et al. Making machine learning models interpretable
- [Ribeiro et al,. 2016] Why Should I Trust You?" Explaining the Predictions of Any Classifier
- [Labarta et al, 26] From Attribution to Action: A Human-Centered Application of Activation Steering

Motivations

Plusieurs réponses

- Renforcer la confiance dans les prédictions du modèle.
- Aider à satisfaire les exigences réglementaires et le processus de certification (e.g. AI Act)
- Révéler les biais ou d'autres effets involontaires d'un modèle.
- Détecter et prévenir les cas d'échecs, intervenir sur le modèle



- Bryce Goodman et al. European union regulations on algorithmic decision-making and a"right to explanation".
- Finale Doshi-Velez et al. Accountability of ai under the law: The role of explanation
- Gabriel Cadamuro et al. Debugging machine learning models
- Alfredo Vellido et al. Making machine learning models interpretable
- [Ribeiro et al,. 2016] Why Should I Trust You?" Explaining the Predictions of Any Classifier
- [Labarta et al, 26] From Attribution to Action: A Human-Centered Application of Activation Steering

Un levier réglementaire – le RGPD

- **Article 22** : protection contre certaines décisions fondées *exclusivement* sur un traitement automatisé produisant des effets significatifs. Exemples typiques : refus de crédit en ligne, tri automatique de candidatures, fixation d'un tarif d'assurance.
- **Articles 13–15** : information sur la logique sous-jacente à ces décisions.
- **Jurisprudence**
  - *SCHUFA* (2023) : un score peut lui-même constituer une décision automatisée s'il joue un rôle déterminant.
  - *Dun & Bradstreet Austria* (2025) : l'information doit rendre compréhensibles la procédure et les principes appliqués (opérateur mobile qui a refusé un contrat à une cliente au motif que sa solvabilité est insuffisante avec une évaluation de crédit réalisée de façon automatisée).

Le RGPD n'établit pas un droit illimité à l'explication de toute décision automatisée.

Omnibus numérique

- Ensemble de propositions législatives de la Commission européenne visant à simplifier le droit du numérique et à alléger les charges administratives des entreprises.
- Article 22 serait supprimé : les décisions ne seraient permises que dans le cadre d'un contrat, d'une autorisation légale ou avec le consentement explicite de la personne.

Un levier réglementaire – l'AI Act

- **Obligations fondées sur le risque** : les exigences dépendent de la catégorie de risque du système et du rôle de l'acteur.
- **Usages à haut risque dans notre secteur** (annexe III) : évaluation de la solvabilité des personnes physiques, et **évaluation des risques et tarification en assurance vie et santé**.
- **Transparence et contrôle** : information des déployeurs, contrôle humain et tenue de registres.
- **Article 86** : un droit conditionnel à l'explication pour certaines décisions fondées sur des systèmes d'IA à haut risque.

La réglementation motive l'XAI, mais n'impose pas d'explication technique pour chaque sortie d'un système d'IA.

Quelques définitions

Interprétabilité (Doshi-Velez & Kim, 2017)

L'interprétabilité est la capacité d'être expliqué ou présenté en des termes compréhensibles par un humain.

Explicabilité (Guidotti et al, 2018)

Une explication est une interface entre un preneur de décision et un humain qui est à la fois une **représentation fidèle** du processus de décision et **compréhensible** par les humains.

Definition of Arrieta et al[a]

[a] https://www.sciencedirect.com/science/article/pii/S1566253519308103

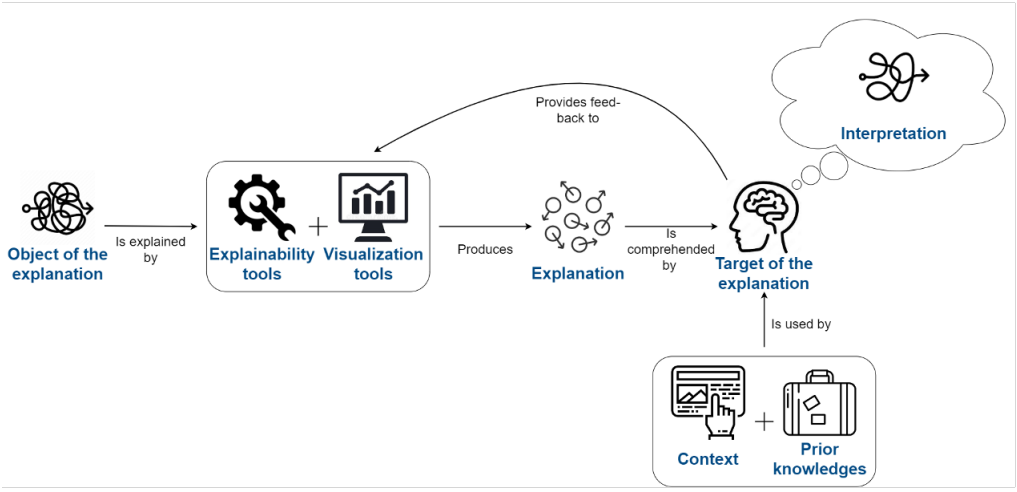
Pour un public donné, l'explicabilité désigne **les détails et les raisons fournis par un modèle** afin de rendre son fonctionnement **clair ou facile à comprendre**.

Interprétabilité vs explicabilité en IA

- On peut considérer que:
  - L'interprétabilité est une **caractéristique et propriété** du modèle, i.e., un humain peut-il comprendre directement comment le modèle passe des entrées aux sorties, à partir de sa structure
  - L'explicabilité nécessite une **procédure pour rendre le modèle plus clair**. On cherche à rendre intelligible une décision ou un comportement, souvent a posteriori.
- Mais très souvent, explicabilité et interprétabilité sont confondues

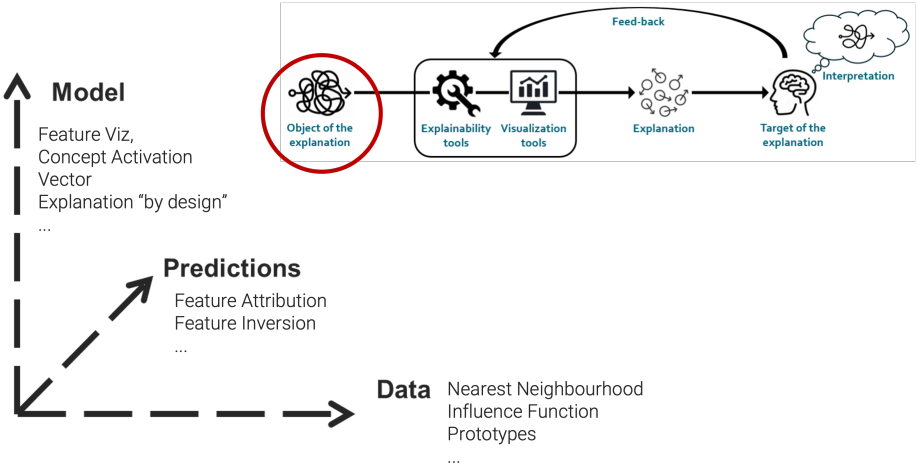
Composantes clés de l'explicabilité

Composantes clés de l'explicabilité



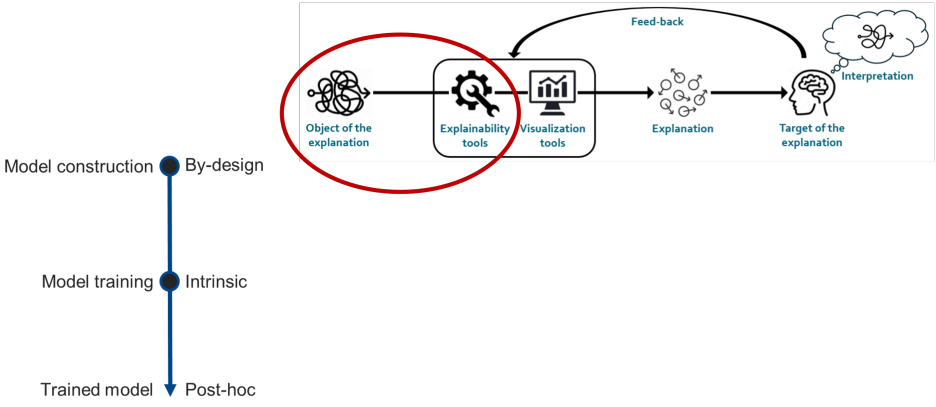
Source: Hudelot, Ouerdane, Fel, Poché – PFIA 2024 tutorial "Some new directions for Explainable AI".

Portée de l'explicabilité



Source: Hudelot, Ouerdane, Fel, Poché – PFIA 2024 tutorial "Some new directions for Explainable AI".

Où ou quand ?



Source: Hudelot, Ouerdane, Fel, Poché – PFIA 2024 tutorial "Some new directions for Explainable AI".

Où ou quand ? (Application time)

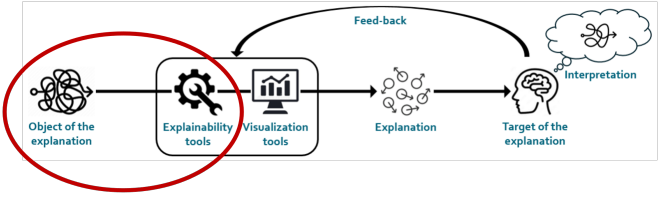
Ante-hoc / Par conception

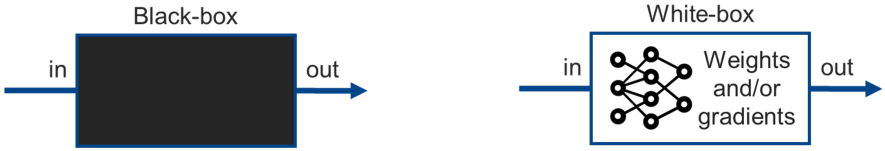
- Interprétabilité interne au modèle : modèles linéaires, arbres, règles, GAM, prototype networks, concept bottlenecks
- L'explication et la prédiction relèvent du même mécanisme.

Post-hoc

- Un modèle entraîné est évalué à l'aide d'une méthode complémentaire.
- LIME, SHAP, gradients, counterfactuals, TCAV...
- La configuration habituelle pour les modèles pré-entraînés.

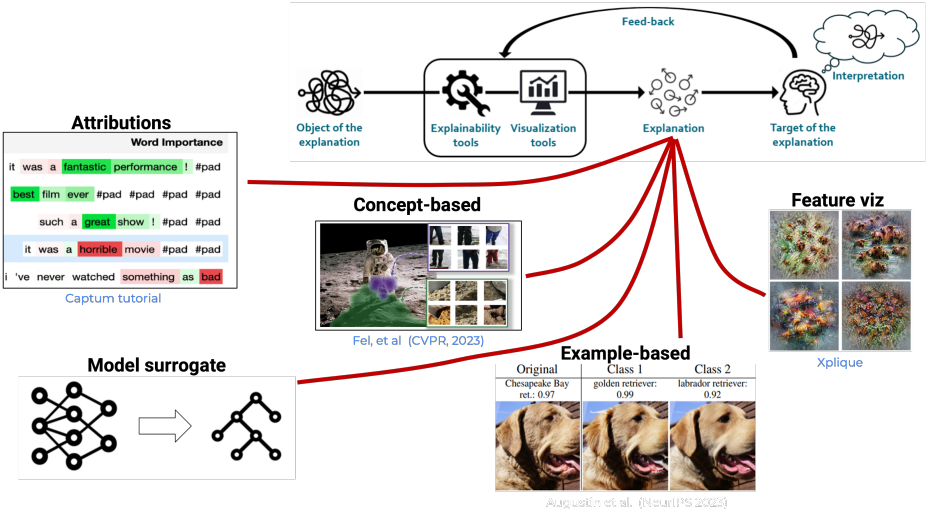
Quelle information est nécessaire pour mettre en place l'explication ?





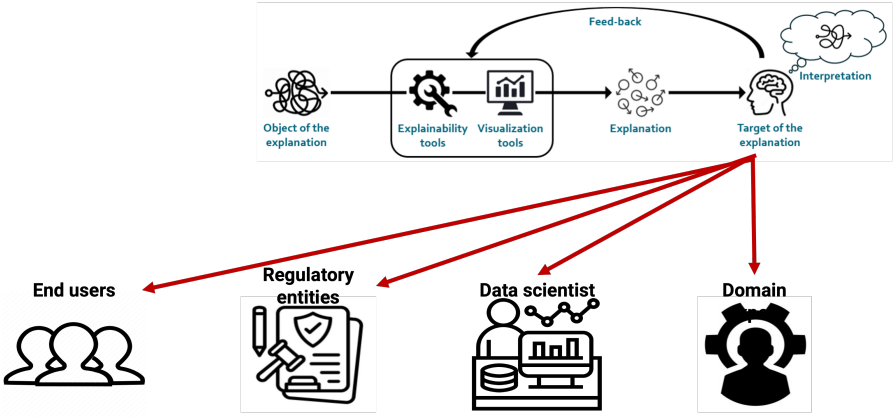
Source: Hudelot, Ouerdane, Fel, Poché – PFIA 2024 tutorial "Some new directions for Explainable AI".

Types d'explication



Source: Hudelot, Ouerdane, Fel, Poché – PFIA 2024 tutorial "Some new directions for Explainable AI".

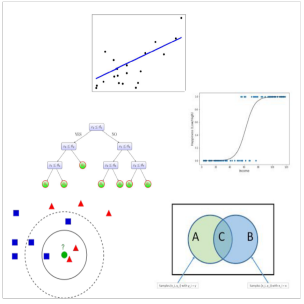
Cibles de l'explication



Source: Hudelot, Ouerdane, Fel, Poché – PFIA 2024 tutorial "Some new directions for Explainable AI".

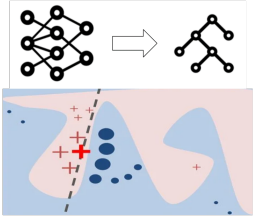
De nombreuses approches

Intrinsic explainable models



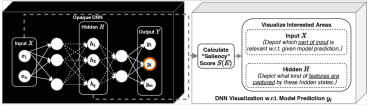
Post-hoc models

Model agnostic

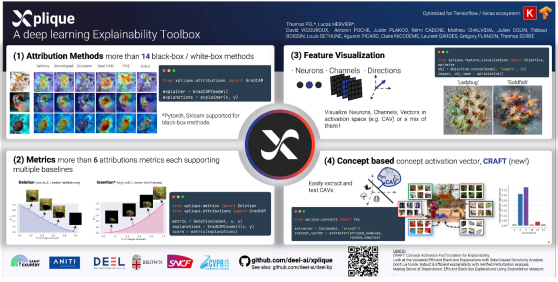


Ribeiro et. al, 2016

Model specific



Des outils





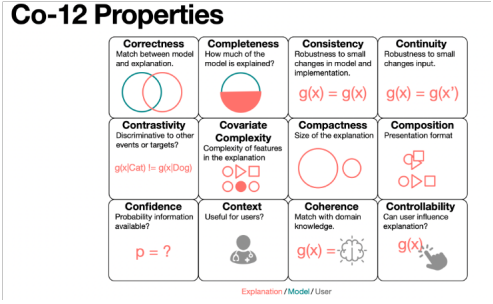
Interpreto

Interpretability Toolbox for LLMs

Comment savoir si une explication est fiable ?

Qu'est-ce qu'une bonne explication ? Une question difficile

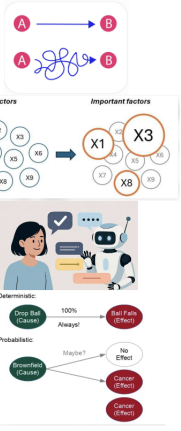
- Plusieurs types et formats sont possibles.
- De nombreux critères, certains **dépendants de l'utilisateur**, certains **antagonistes** (fidèle vs. simple).
- Pas d'explication de référence (vérité terrain), pas de benchmark universel.



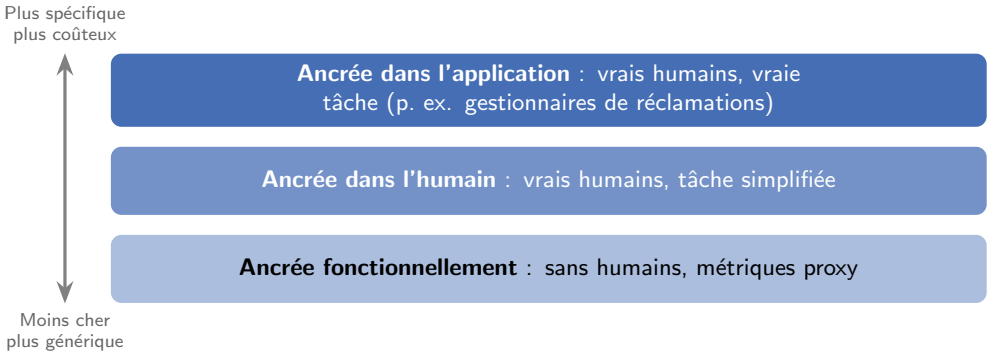
Source : Nauta et al., ACM Computing Surveys 2023

Qu'est-ce qu'une bonne explication pour les humains ? (Miller, 2019)

- **Contrastive** : ≪ pourquoi P *plutôt que* Q ? ≫
- **Sélective** : quelques causes nécessaires et suffisantes, pas toutes (risque de biais cognitif dans la sélection).
- **Sociale** : s'inscrit dans un dialogue ; dépend des croyances de celui qui explique et de celui qui reçoit l'explication.
- **Des causes plutôt que des probabilités** : les associations statistiques satisfont moins que les causes.



Trois niveaux d'évaluation (Doshi-Velez & Kim, 2017)



La plupart des articles s'arrêtent aux métriques fonctionnelles ; les décisions de déploiement exigent des preuves ancrées dans l'humain.

Fidélité, plausibilité, confiance – trois propriétés différentes

Fidélité

L'explication reflète-t-elle le **raisonnement réel** du modèle ?

Sans humains : suppression / insertion, tests de randomisation.

Plausibilité

Est-elle **convaincante pour les humains** ? Correspond-elle aux justifications humaines ?

Avec humains : annotations de justifications (p. ex. ERASER), études utilisateurs.

Confiance

Conduit-elle à une **confiance appropriée** : se fier au modèle quand il a raison, s'en méfier quand il a tort ?

Études utilisateurs *comportementales*.

Une explication plausible n'est pas nécessairement fidèle – avoir du sens pour un humain ne signifie pas refléter les indices utilisés par le modèle. Source : Jacovi & Goldberg, ACL 2020 ; Jacovi et al., FAccT 2021 ; DeYoung et al., ACL 2020.

Exemple : une explication convaincante. . .

Une réclamation est routée vers **transaction non autorisée**. Le gestionnaire voit une **carte d'attention**, une explication courante :

≪ Je n'ai jamais autorisé ce paiement de 450 €, merci de bloquer ma carte. ≫

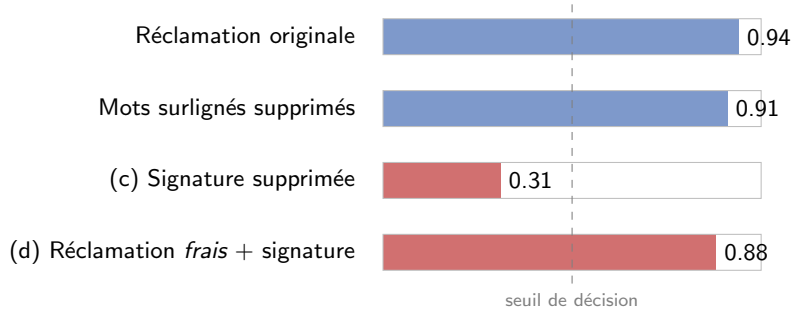
– Envoyé depuis l'agence Riverside

⇒ les mots qu'un expert choisirait : la carte est **plausible**.

Mais : dans les données d'entraînement, **9 sur 10** des réclamations ≪ transaction non autorisée ≫ étaient signées ≪ *Envoyé depuis l'agence Riverside* ≫.

. . . qui cache un raccourci

On modifie l'entrée et on observe le score de *transaction non autorisée* : *(illustratif)*



(c) et (d) sont aussi des explications – elles **modifient l'entrée** – et elles contredisent la carte : le score suit la **signature**. La carte était **plausible** mais **pas fidèle**.

Les pièges côté humain

Les explications augmentent l'acceptation – y compris quand le modèle se trompe

Sur-confiance

Une carte d'attributiion convaincante pousse le gestionnaire à accepter le routage – correct *ou* erroné. Les explications n'aident pas toujours les utilisateurs à distinguer les prédictions correctes des incorrectes (Nguyen et al., 2021 ; Hase & Bansal, 2020).

Biais de confirmation

≪ *jamais autorisé* ≫ est surligné, comme attendu ⇒ l'expert cesse de chercher et ne voit jamais le rôle de la **signature**.

Où ≠ quoi

≪ *paiement* ≫ est surligné – mais quelle notion : le montant, la contestation, l'urgence ? ≪ Montrer où un réseau regarde ne nous dit pas ce qu'il voit ≫ (Achtibat et al., 2023)

Messages à retenir

- Une prédiction correcte ne suffit pas : un modèle peut avoir **raison pour de mauvaises raisons**.
- Décrire toute explication par son **objet**, son **accès**, sa **forme** et son **destinataire**.
- **Plausible** $\neq$ **fidèle** $\neq$ **utile**
- Une explication est une **information à évaluer**, pas une garantie.

Une vue rapide sur les méthodes d'attributions - Cas du texte

Pourquoi expliquer les modèles de langue est difficile ?

Modèles de langues

- **Représentation du texte** : le texte est découpé en **tokens**, chacun associé à un **plongement** (*embedding*) de grande dimension.
- **Architecture** : presque tous les modèles récents sont des **transformers**, construits autour de l'attention.
- **Tâches** : classification et **génération** – traduction, question-réponse, résumé, assistants.

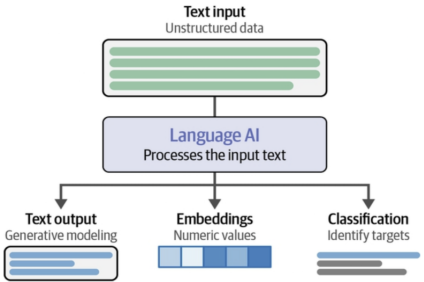
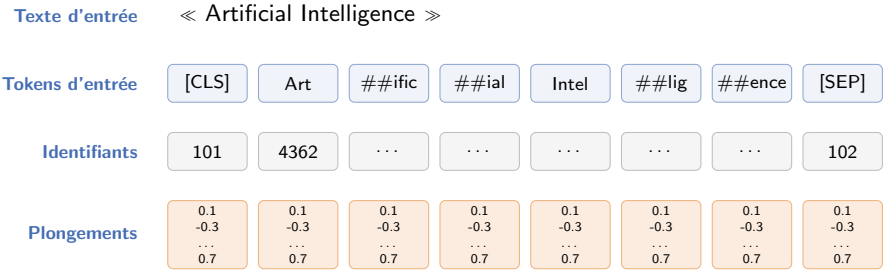


Figure: Source : Alammar and Grotendorst - Book 2024 - Hands-On Large Language Models

Défi 1 – Tokenisation et plongements



- Le modèle ne voit jamais de ≪ mots ≫ : il voit des **tokens**, souvent des sous-mots, puis des **vecteurs** de centaines de dimensions.
- Une explication par token doit être **ré-agrégée** en mots pour être lisible – un choix qui change le résultat.

Défi 2 – Charge cognitive

≪ *Bonjour, le 3 mars j'ai constaté sur mon relevé un paiement par carte de 450 € dans un magasin où je ne suis jamais allé, alors que ma carte était dans mon portefeuille ; j'ai appelé l'assistance le jour même, on m'a dit de vous écrire, ce que j'ai fait le 5, mais je n'ai eu aucune réponse depuis, et entre-temps on m'a prélevé 12 € de frais pour un découvert non autorisé causé par ce paiement. . .* ≫

- des centaines de tokens, chacun avec un score : **qui lit ça** ?
- l'explication doit être **résumée** – par phrase, par segment, par concept ;
- et un résumé est déjà une **interprétation**.

Défi 3 – Contexte et ambiguïté

Un même mot, plusieurs sens – seul le **contexte** tranche :

- « On m'a **prélevé** deux fois » → des frais, ou un débit ?
- « ma **carte** a été **bloquée** » → protection anti-fraude, ou retard ?
- « Je n'ai **jamais** reçu le virement, **pas** le paiement » → portée de la négation.

- La représentation d'un mot **dépend de ses voisins** : son importance ne peut pas être lue isolément.
- Les mots **interagissent** (négation, portée) : en retirer un change le sens des autres.

Défi 4 – Génération

- Le texte est généré **token par token** : chaque nouveau token est prédit à partir du prompt *et* des tokens déjà générés (autorégressif).
- Une explication locale couvre **un seul** token prédit.
- Expliquer une réponse entière nécessite **autant d'explications que de tokens générés** – puis un moyen de les résumer.



Source : Sarti et al. - ACL 2023 - Inseq: An Interpretability Toolkit for Sequence Generation Models

Défi 5 – Auto-explications

Une bonne explication doit être (Jacovi & Goldberg, 2020) :

- **plausible** : convaincante pour les humains ;
- **fidèle** : refléter le mécanisme qui a réellement produit la sortie.

Défi 5 – Auto-explications

Une bonne explication doit être (Jacovi & Goldberg, 2020) :

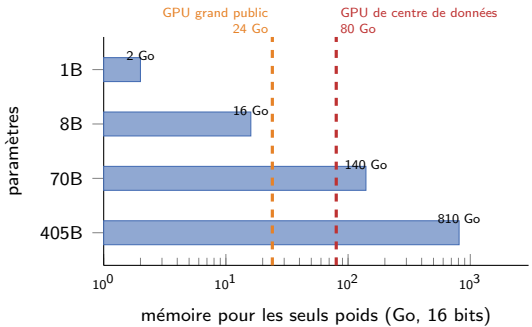
- **plausible** : convaincante pour les humains ;
- **fidèle** : refléter le mécanisme qui a réellement produit la sortie.

Les LLM peuvent **s'expliquer eux-mêmes** en langage naturel :

- ces explications sont très **plausibles** – les modèles sont entraînés à produire du texte convaincant ;
- rien ne montre qu'elles sont **fidèles** (Barez et al., 2025).

Défi 6 – Taille des modèles

- Une explication nécessite souvent **des centaines d'inférences** (perturbations, chemins, échantillons).
- Un modèle 16 bits de plus de ∼14 milliards de paramètres ne tient pas sur un **GPU grand public** (32 Go) ; la quantification aide l'inférence, mais les explications nécessitent **des gradients et de nombreuses passes**, et quantifier change le modèle expliqué.
- Les plus grands sont **répartis sur plusieurs GPU** : exécuter une méthode d'explication devient un défi d'**ingénierie**.



≈ 2 octets par paramètre ; les activations et les gradients (pour les méthodes à gradient) s'y ajoutent.

Méthodes d'attribution : définition

Définition (Méthode d'attribution). *Pour un modèle* $\boldsymbol{f} : \mathcal{X} \to \mathcal{Y}$ *et une entrée* $\boldsymbol{x} \in \mathcal{X}$*, une méthode d'attribution est une fonctionnelle :*

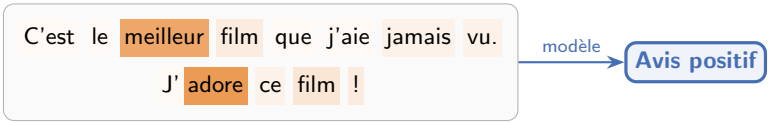
$$\boldsymbol{\Phi} : \mathfrak{F} \times \mathcal{X} \to \mathbb{R}^{|\mathcal{X}|}$$

où $\boldsymbol{\gamma} = \boldsymbol{\Phi}(\boldsymbol{f}, \boldsymbol{x})$ *(avec* $\boldsymbol{f} \in \mathfrak{F}$*) représente une carte d'attribution qui explique la prédiction de* $\boldsymbol{f}$ *pour l'entrée* $\boldsymbol{x}$*. Plus la valeur scalaire dans* $\boldsymbol{\gamma}$ *est élevée, plus la variable est considérée comme importante.*

Source : T. Fel, thèse de doctorat (2024)

Attribution pour la classification

Tâche de classification – quels mots poussent l'avis vers la classe prédite ?



Carte de chaleur de l'importance des mots pour la classe ≪ **positif** ≫

importance faible → forte

Adapté du tutoriel XAI PFIA 2025

Attribution : application à la détection de biais

Tâche de détection de biais – un classifieur de professions entraîné sur des biographies en français.

Elle travaille à l'hôpital de Perpignan. → Classe prédite : **Infirmière**
Classe réelle : **Chirurgienne**

Pourquoi *infirmière* ? Regardons les mots qui conduisent à la **mauvaise** classe.

Attribution : application à la détection de biais

Tâche de détection de biais – un classifieur de professions entraîné sur des biographies en français.

Elle travaille à l' hôpital de Perpignan. →

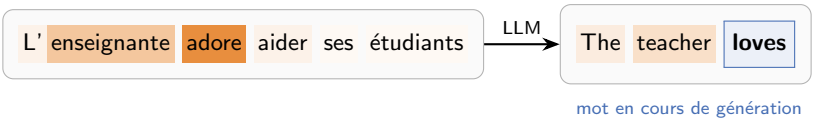
Classe prédite : **Infirmière**

Classe réelle : **Chirurgienne**

Carte de chaleur de l'importance des mots pour la prédiction de la classe ≪ **infirmière** ≫ *(scores illustratifs)*

XAI par attribution pour la génération

Tâche de génération – traduction : quels mots précédents comptent pour le *prochain* mot généré ?



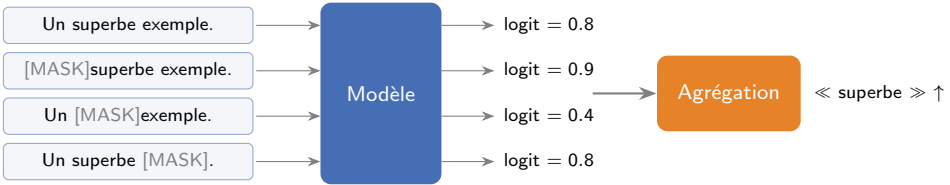
En génération, il y a **une carte d'attribution par token généré**, sur le prompt *et* les tokens déjà générés.

Une vue unifiée : trois unités, trois techniques

	Perturbation	Gradients	Approximation linéaire
Caractéristiques d'entrée quelles parties de l'entrée ?	Occlusion, RISE, Sobol, valeurs de Shapley	Saillance, Grad×Entrée, SmoothGrad, IG, LRP	LIME, KernelSHAP
Données d'entraînement quels exemples d'entraînement ?	Leave-one-out, Data Shapley	Fonctions d'influence, TracIn	Datamodels, TRAK
Composants du modèle quelles unités internes ?	Ablation, *activation patching*, *causal tracing*	*Attribution patching*	COAR

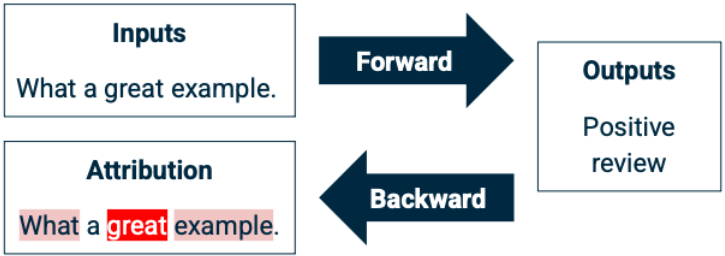
Zhang, Han, Bhalla & Lakkaraju, 2025 (arXiv:2501.18887)

Attribution par perturbation : le principe



Deux questions définissent toute méthode par perturbation : **comment perturber** l'entrée ? **comment agréger** les scores ?

Attribution par gradient : le principe



$$\Phi = \nabla f(x) \implies \phi_i = \frac{\partial f(x)}{\partial x_i}$$

Where do we compute the gradient?

How do we aggregate gradients?

Approximations linéaires : perturber, puis ajuster un modèle linéaire

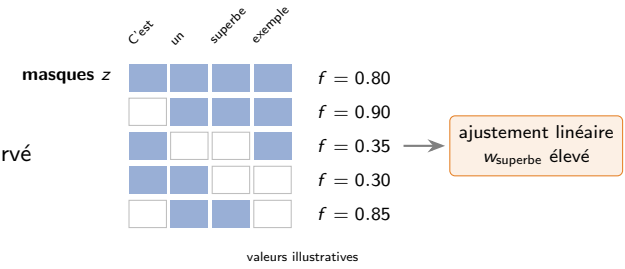
Mêmes requêtes que les méthodes par perturbation : **des textes dont des mots sont masqués aléatoirement**.

Agrégation différente : ajuster un modèle linéaire sur les paires (masque, score)

$f(x_z) \approx w_0 + \sum_i w_i z_i, \qquad z_i = 1$ si le mot i est conservé

l'attribution du mot i est son **coefficient** w_i.

Deux choix : **quels échantillons comptent le plus** (le noyau) et le degré de parcimonie du modèle.



Perturbation et approximation linéaire posent au modèle **le même type de questions** ; elles diffèrent par **la façon de résumer les réponses**.

Sources : Covert, Lundberg & Lee, JMLR 2021 (≪ explaining by removing ≫) ; Zhang et al., 2025.

Explicabilité par concepts

Pourquoi expliquer avec des concepts ?

Pourquoi des concepts ? (1) Les attributions sont fragiles

Une infime modification de l'entrée – même sens, même prédiction – peut donner une **carte très différente**.

(illustratif)

Texte	**Carte pour transaction non autorisée**	**Prédiction**
original	n'ai jamais autorisé ce paiement de 450 €	non autorisée (0,94)
un mot changé	n'ai jamais approuvé ce paiement de 450 €	non autorisée (0,93)

- De petites perturbations qui préservent le sens peuvent beaucoup déplacer la carte (Ghorbani, Abid & Zou, 2019).
- Les cartes peuvent même être **manipulées** sans que les prédictions changent (Anders et al., 2020).

Pourquoi des concepts ? (2) Quelle classe explique-t-on ?

Deux cartes de la **même réclamation**, pour deux classes différentes *(illustratif)* :

non autorisée (0,94)	Je n'ai jamais autorisé ce paiement de 450 €, on m'a débité deux fois
frais (0,04)	Je n'ai jamais autorisé ce paiement de 450 €, on m'a débité deux fois

- Les deux cartes se ressemblent, et **toutes deux semblent plausibles**.
- **Biais de confirmation** : une carte qui nous paraît sensée ne prouve rien sur la prédiction (Rudin, 2019).

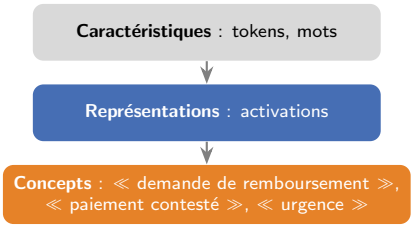
Toujours préciser **quelle classe** est expliquée – et ne pas prendre « ça a du sens » pour un test.

Pourquoi des concepts ? (3) Du *où* au *quoi* ?

Une carte de saillance montre **où** le modèle regarde, pas **ce qu'il** y voit (Rudin, 2019 ; Achtibat et al., 2023).

Surligner paiement ne dit pas *quelle notion* le modèle utilise :

- le **montant** ?
- la **contestation** ?
- l'**urgence** ?



Expliquer avec des **notions de haut niveau** partagées par de nombreux textes – pas avec des tokens.

Explications par concepts : l'objectif

Objectif – Expliquer avec des **concepts humains de haut niveau** plutôt qu'en attribuant une importance à chaque token. Les concepts sont des **propriétés qui correspondent à l'intuition humaine** : ≪ paiement contesté ≫, ≪ frais ≫, ≪ urgence ≫.

Propriétés souhaitées d'un concept (Ghorbani et al., 2019)

- **Signifiant** : un exemple du concept a un sens en lui-même – une expression, pas un sous-token.
- **Cohérent** : ses exemples se ressemblent entre eux et diffèrent des exemples des autres concepts.
- **Important** : sa présence est nécessaire à la bonne prédiction de la classe.

Source Ghorbani et al., NeurIPS 2019.

Motivation 1 – Plus proche du raisonnement humain

Les utilisateurs préfèrent les explications en termes de **concepts** aux cartes de chaleur (Kim et al., 2023).

Un chargé de réclamations raisonne avec des notions – « paiement contesté », « frais », « signature d'agence » – pas avec des scores de tokens.

Source : Kim, S. S. Y. et al., CHI 2023.

Carte de tokens

Je n'ai jamais autorisé ce paiement de 450 € – Envoyé depuis l'agence Riverside

Explication par concepts

paiement contesté	0,35
signature d'agence	**0,55**
montant	0,05

valeurs illustratives

Motivation 2 – Plus stable

Des entrées similaires devraient recevoir des **explications similaires** (Alvarez-Melis & Jaakkola, 2018). *(illustratif)*

Texte	**Carte de tokens**	≪ **paiement contesté** ≫
Je **n'ai jamais autorisé** ce paiement	n'ai jamais autorisé ce paiement	0,81
Je **n'ai pas approuvé** ce paiement	n'ai pas approuvé ce paiement	0,79

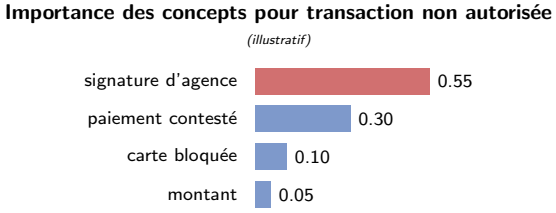
Une paraphrase déplace beaucoup la carte de tokens ; le score du concept bouge à peine : le concept **regroupe** les nombreuses façons de dire la même chose. Source : Alvarez-Melis & Jaakkola, NeurIPS 2018.

Motivation 3 – Mieux révéler les biais

Un biais réparti sur de nombreux tokens et de nombreux textes est difficile à voir dans les cartes de tokens.

Sous forme de **concept**, il reçoit un nom et un score testable sur tout le jeu de données : ≪ genre féminin ≫ pour un classifieur de professions, ≪ signature d'agence ≫ pour des réclamations (Jain et al., 2022).

Source : Jain et al., EMNLP 2022.



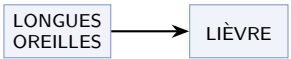
Explications par concepts : principes généraux

Concepts : différents types

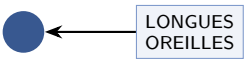
- **Concepts symboliques** : symboles définis par l'humain – attributs de haut niveau de la tâche ou abstractions interprétables.
- **Bases de concepts non supervisées** : clusters ou directions appris par le modèle – des abstractions plus compréhensibles que les tokens individuels.
- **Prototypes** : exemples représentatifs de traits particuliers des données d'entraînement.
- **Concepts textuels** : descriptions textuelles des principales classes – de plus en plus présents avec les LLM.

Source : Poeta et al., arXiv:2312.12936

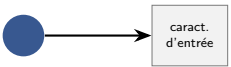
Explications par concepts : différents types



Relation classe–concept
relation entre un concept et une classe de sortie du modèle : importance du concept pour la classe, ou règle logique sur plusieurs concepts



Association nœud–concept
association explicite d'un concept à des nœuds cachés du réseau



Visualisation de concept
visualisation d'un concept appris en termes de caractéristiques d'entrée – pour du texte, les mots ou extraits qui le représentent le mieux

XAI par concepts pour la classification

*Exemple 1 – **une autre tâche** (classification de professions) : quels concepts, où dans le texte, avec quel poids.*

Tâche de classification – prédire une profession à partir d'une biographie en français.

Le Docteur Lecomte travaille au service d'odontologie de l'Hôpital Saint-Martin. Diplômé de la Faculté de Chirurgie Dentaire de Paris, il s'est spécialisé en soins dentaires hospitaliers, notamment dans la prise en charge des patients souffrant de pathologies complexes.



Classe prédite : **Dentiste**

72 %	Concept 4	*études supérieures*
63 %	Concept 12	*médical*
87 %	Concept 32	*dentaire*

XAI par concepts : application à la détection de biais

Exemple 2 – même classifieur de professions : un biais devient un concept nommé.

Tâche de détection de biais – même classifieur de professions.

Elle travaille à l'hôpital de Perpignan depuis 3 ans. Les patients qu'elle opère la recommandent fortement : son sérieux et sa gentillesse.



Prédit : **Infirmière**
Vrai : **Chirurgienne**

70 %	Concept 12	*médical*
90 %	Concept 17	*genre féminin*

Un texte est une **hypothèse**, pas une preuve : la tester sur de nombreuses biographies en ne changeant que les marqueurs de genre (*il/elle*, prénoms) et en comparant les prédictions.

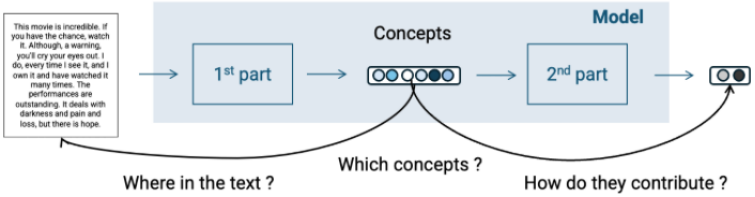
Lire une explication par concepts : locale et globale

Locale – pour un texte

1. **quels** concepts sont présents ?
2. **où** sont-ils dans le texte ?
3. **comment** contribuent-ils ?

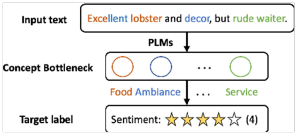
Globale – pour le modèle

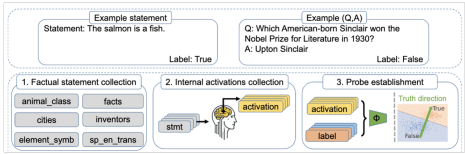
- le lien entre **concepts** et **classes** : ≪ le concept 3 pilote *frais* ≫.

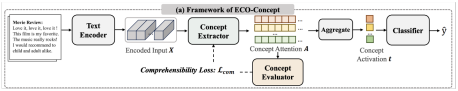


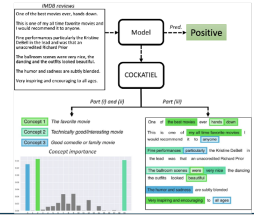
Explication par concepts : différentes approches

	Ante-hoc The model is trained to reason from concepts	**Post-hoc** Concepts are identified within the trained model
Supervised Requires labelled concepts	e.g. Tan, Zhen et al. - PACKDDM 2024	e.g. Bao, Yuntai, et al. - 2025
Unsupervised Annotation free	e.g. Sun, Yifan, et al. - ACL 2025	e.g. Jourdan, Fanny, et al. - ACL 2023



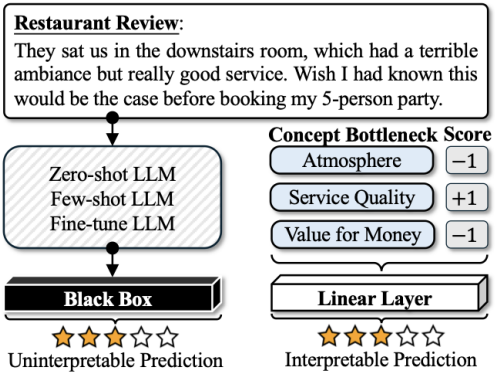






Supervisé ante-hoc : Concept Bottleneck Models

- Entrée : un texte x et une étiquette cible y.
- Le modèle prédit des concepts interprétables : g : texte $\rightarrow$ concepts.
- Un classifieur f prédit l'étiquette à partir de ces concepts.
- Les concepts fournissent une représentation intermédiaire interprétable de la prédiction.



Source: Ludan et al., 2023, Text Bottleneck Models

Concept Bottleneck Models : trois schémas d'entraînement

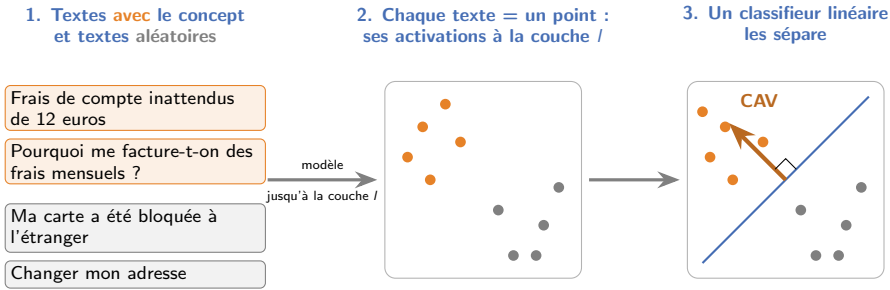
Un Concept Bottleneck Model prédit une étiquette en passant par des concepts définis par l'humain : $x \xrightarrow{g} \hat{c} \xrightarrow{f} \hat{y}$ où g prédit les concepts et f prédit l'étiquette à partir de ceux-ci.

Schéma	**Mode d'entraînement**	**Concepts utilisés par f**
Indépendant	Entraîner g et f séparément	Concepts réels c
Séquentiel	Entraîner g, puis entraîner f sur ses prédictions	Concepts prédits $\hat{c}$
Conjoint	Entraîner g et f ensemble	Concepts prédits $\hat{c}$

Source : Koh et al., ICML 2020.

Supervisé Post-hoc : TCAV

Question : le modèle entraîné possède-t-il une notion interne de *frais* ?

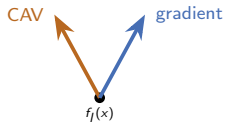


Source : (Kim et al., 2018)

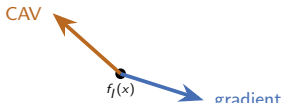
Concept Activation Vector : sensibilité conceptuelle

Question : pour *cette* réclamation x, « plus de concept C » fait-il **monter** ou **baisser** le score de la classe k ?

$$S_{C,k,l}(x) = \underbrace{\nabla h_{l,k}\big(f_l(x)\big)}_{\text{direction qui augmente la classe } k} \cdot \underbrace{v_C^l}_{\text{CAV}}$$



même direction : $S > 0$
le concept augmente la classe k



directions opposées : $S < 0$
le concept diminue la classe k

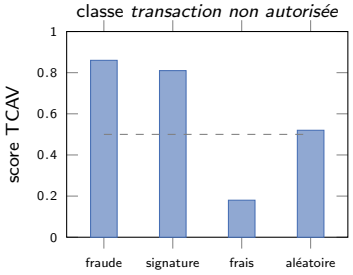
Exemple : C = *signature d'agence*, k = *transaction non autorisée*. Même idée que la saillance, mais le gradient est lu **le long d'un concept** au lieu de l'être le long de chaque token. Source : Kim et al., ICML 2018.

D'un texte à une classe : le score TCAV

Pour la classe k, avec X_k les entrées de cette classe :

$$\mathrm{TCAV}_{C,k,l} = \frac{\left|\{x \in X_k : S_{C,k,l}(x) > 0\}\right|}{|X_k|}$$

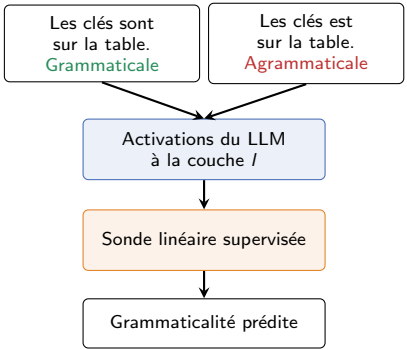
- proportion des textes de la classe k dont le score **augmente** quand on se déplace vers C ;
- ne dépend que du **signe** de S – des variantes utilisent sa magnitude ;
- une explication **globale** : un nombre par (concept, classe, couche).



Supervisé post-hoc – sondage (probing)

Une **sonde** est un classifieur léger entraîné à prédire une propriété linguistique à partir des représentations internes d'un modèle de langue gelé.

1. Des phrases étiquetées selon une propriété linguistique, p. ex. la grammaticalité ;
2. Les **activations** du LLM à la couche l ;
3. Une **sonde linéaire** supervisée entraînée à prédire les étiquettes.

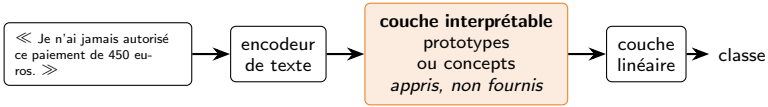


Point clé : une sonde teste quelle information est **décodable** à partir des représentations, pas nécessairement ce que le modèle **utilise** (Belinkov, 2022).

Source : Belinkov, Computational Linguistics (2022)

Non supervisé ante-hoc : le principe

Idée : aucune annotation de concepts – une **couche interprétable** à l'intérieur du modèle apprend ses propres concepts, et la décision est calculée **à partir** d'eux.



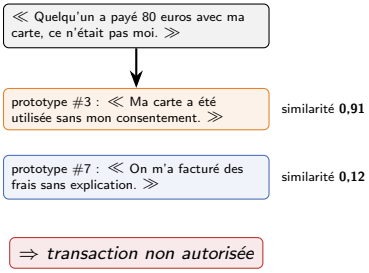
l'explication est le calcul

- **Prototypes** : ≪ ce texte ressemble à celui-là ≫ – ProtoPNet en vision, ProSeNet pour le texte.
- **Concepts** : concepts découverts à partir du texte – ECO-Concept.

Prototypes pour le texte – ProSeNet

ProSeNet (Ming et al., 2019)

- le texte est encodé et comparé à quelques **séquences prototypes** ;
- les prototypes sont de **vrais textes d'entraînement** ;
- classe = somme pondérée des **similarités** ;
- les experts peuvent **modifier** les prototypes.



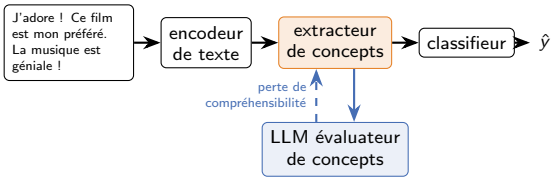
Exemple illustratif.

Source : Ming et al., *Interpretable and steerable sequence learning via prototypes*, KDD 2019.
Équivalent en vision : ProtoPNet (Chen et al., 2019).

Concepts pour le texte – ECO-Concept

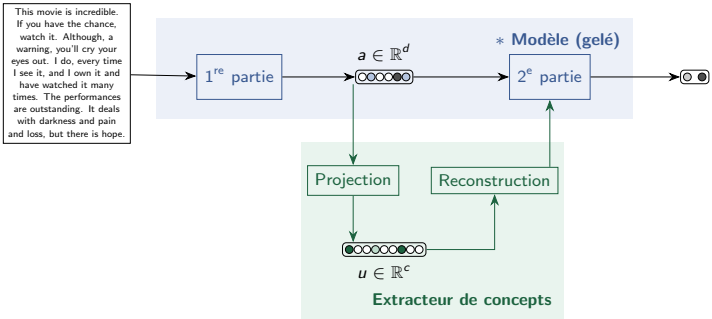
ECO-Concept (Sun et al., 2025)

- un **extracteur de concepts** apprend des emplacements de concepts à partir du texte ;
- la classe est prédite à partir des **concepts** ;
- un **LLM** évalue la compréhensibilité de chaque concept – ce score est une **fonction de perte** d'entraînement.



Source : Sun, Wang, Sheng, Cao & Li, arXiv:2505.20293 (2025)

Un cadre pour la C-XAI post-hoc non supervisée



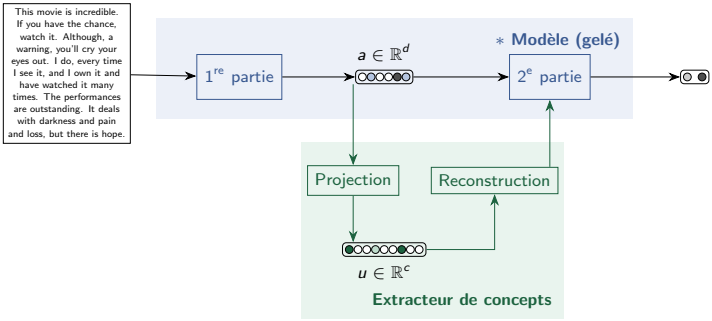
Un cadre pour la C-XAI post-hoc non supervisée

② Choisir les échantillons

Sélection d'échantillons (p. ex. par classe) [1]
Parties d'échantillons (p. ex. phrases)

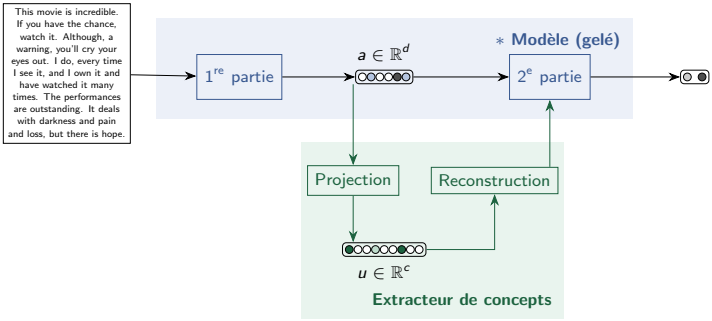
① Choisir une couche

Plongements de tokens / de textes [2] / [1]
Multi-couches [3]
Concepts de bas / haut niveau [4]



Un cadre pour la C-XAI post-hoc non supervisée

② **Choisir les échantillons** ① **Choisir une couche**



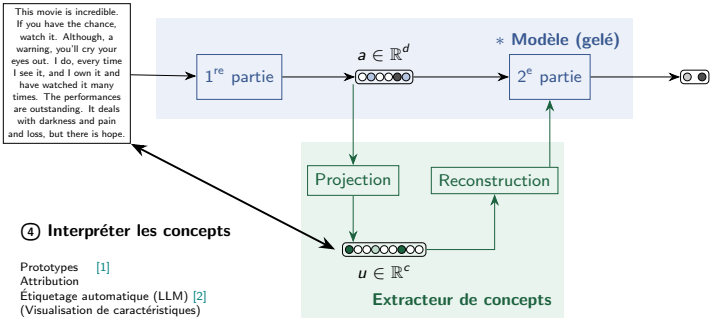
③ **Choisir et entraîner un extracteur de concepts**

Apprentissage de dictionnaire (NMF, SAE, ...) NMF [1], SAE [2]
Concepts 1D / multidimensionnels
Concepts linéaires / non linéaires
Concepts hiérarchiques [3]
Nombre de concepts : surcomplet / sous-complet [2] / [1]

Un cadre pour la C-XAI post-hoc non supervisée

② **Choisir les échantillons**

① **Choisir une couche**

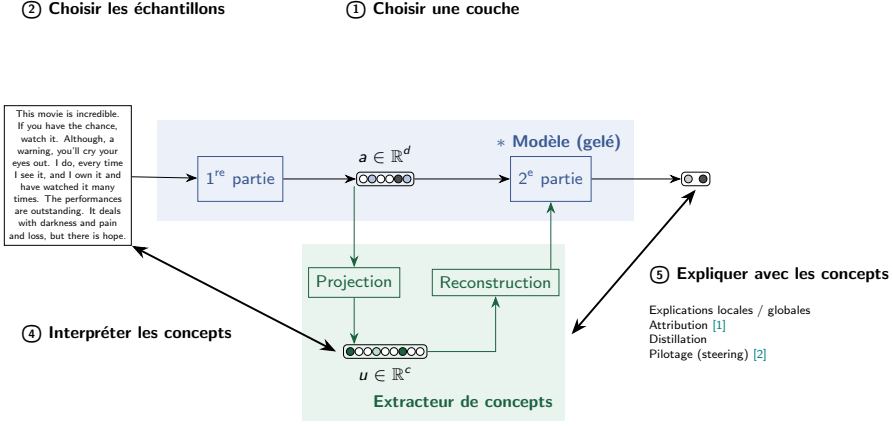


④ **Interpréter les concepts**

Prototypes [1]
Attribution
Étiquetage automatique (LLM) [2]
(Visualisation de caractéristiques)

③ **Choisir et entraîner un extracteur de concepts**

Un cadre pour la C-XAI post-hoc non supervisée

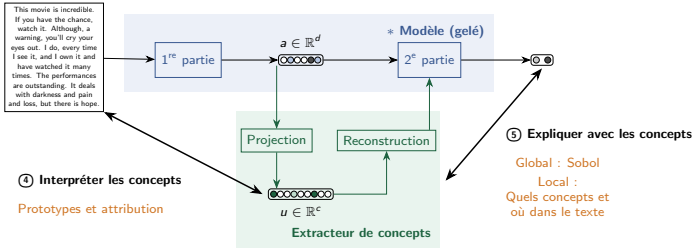


COCKATIEL

② **Choisir les échantillons**

Phrases / propositions

① **Choisir une couche** Couche cachée du classifieur



④ **Interpréter les concepts**

Prototypes et attribution

⑤ **Expliquer avec les concepts**

Global : Sobol

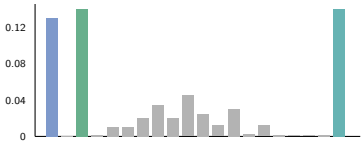
Local :
Quels concepts et où dans le texte

③ **Choisir et entraîner un extracteur de concepts**

Factorisation en matrices non négatives (NMF)

Concept 1	*Le film préféré*
Concept 2	*Film techniquement bon/intéressant*
Concept 3	*Bonne comédie ou film familial*

Importance des concepts



One of the best movies ever hands down

This is one of my all time favorite movies I would recommend it to anyone

Fine performances particularly the Kristine DeBell in the lead was that an unaccredited Richard Prior

The ballroom scenes were very nice the dancing the outfits looked beautiful

The humor and sadness are subtly blended

Very inspiring and encouraging to all ages

Messages à retenir

- Les concepts font passer les explications du **où** au **quoi**.
- Distinguer trois niveaux : **défini par l'humain**, **opérationnalisé**, **interprété**.
- **Lisible** dans les activations ne veut pas dire **utilisé** ; un test **comportemental** montre une influence, pas une représentation.
- Les goulots ne sont fidèles qu'en l'absence de **fuite** ; les concepts extraits (NMF, SAE) sont des **choix**, ni uniques ni stables par défaut.
- Les étiquettes de concepts doivent être **validées avec des experts**.

Transparence des systèmes d'IA

Propriété du système d'IA et de son cycle de vie

Rendre disponibles les informations sur le système : données d'entraînement, finalité, architecture, performances et limites, conditions d'usage,

4 dimensions de transparence

1. Transparence des Données

Disclosure explicite des sources, méthodes de nettoyage, copyrights, filtrage et conditions de travail des annotateurs humains.

2. Transparence Algorithmique

Accès à l'architecture du modèle, hyperparamètres, poids (Open Weights) et logs de calculs de l'entraînement.

3. Transparence Explicative

Capacité d'interpréter pourquoi une décision spécifique a été prise pour un utilisateur final ou un auditeur régulateur (XAI).

4. Transparence d'Impact

Publication des métriques d'évaluation, biais mesurés, rapports d'incidents post-déploiement et empreinte carbone.

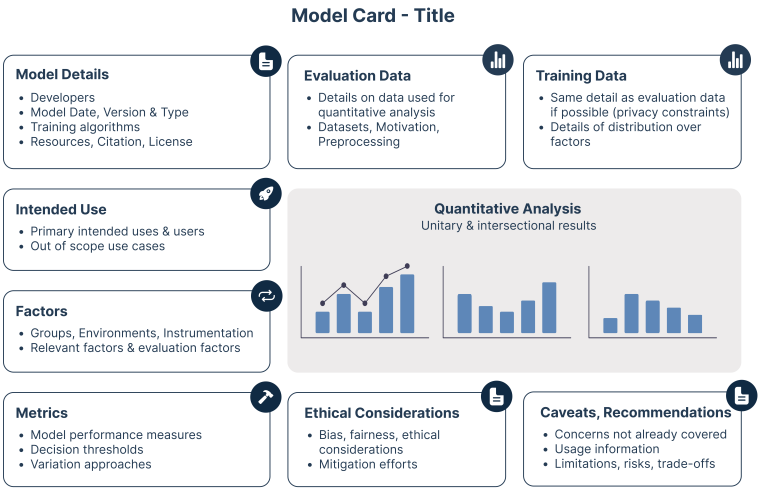
EU AI Act : La Transparence est une Obligation Légale

Les outils pour déclarer

Artefact	Origine	Ce qu'il documente
Model Cards	Mitchell et al., 2019	Usage prévu, performances désagrégées, limites
Datasheets	Gebru et al., 2018	Motivation, collecte, composition, maintenance d'un jeu de données
Data Statements	Bender & Friedman, 2018	Caractéristiques linguistiques et démographiques (NLP)
FactSheets	IBM, Arnold et al., 2019	Déclaration de conformité d'un service d'IA
Data Cards	Google, Pushkarna et al., 2022	Version structurée et modulaire des datasheets
System Cards	OpenAI, Anthropic, Meta…	Système déployé : évaluations de risques, red teaming, mitigations

Model card

9 sections proposées par Mitchell et al. (FAT* 2019)



Autres artefacts de transparence

SYSTEM CARD

GPT-4 System Card · OpenAI, 2023

Red teaming externe, risques et mitigations. Le rapport technique, lui, ne donne ni architecture ni données.

https://cdn.openai.com/papers/gpt-4-system-card.pdf

SYSTEM CARD

Claude system cards · Anthropic

Évaluations de capacités dangereuses selon la Responsible Scaling Policy, tests d'alignement.

https://www.anthropic.com/system-cards

DATASHEET

The Pile · Biderman et al., 2022

Datasheet rétrospective d'un corpus de pré-entraînement de 825 Go, source par source.

https://arxiv.org/pdf/2201.07311

RÉGLEMENTAIRE

AI Act · modèles GPAI

Documentation technique (annexes XI–XII), résumé public des données d'entraînement, Code of Practice.

Une méthodologie

FMTI : Foundation Model Transparency Index - Stanford

1 Concevoir les indicateurs

2 Choisir les entreprises

3 Collecter l'information (publique + rapports)

4 Noter chaque indicateur 0 ou 1

5 Dialoguer avec les entreprises

UPSTREAM
34
Données, travail humain, compute, énergie

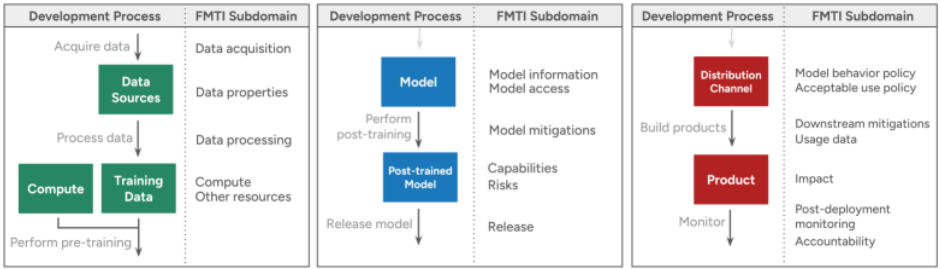
MODEL
30
Propriétés, accès, capacités, risques, mitigations

DOWNSTREAM
36
Usage, impact, monitoring, politiques

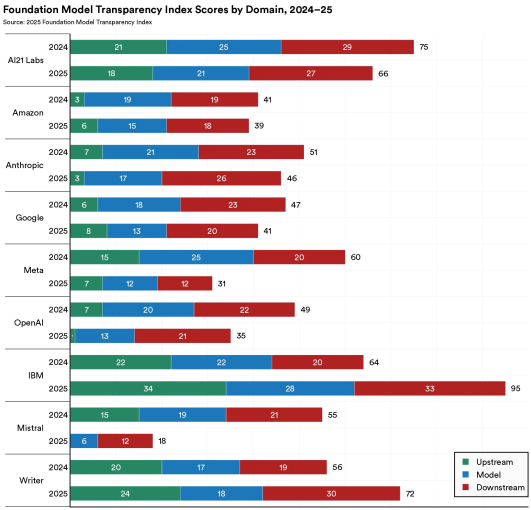
100 indicateurs binaires, pondération égale, 18 sous-domaines (édition 2025) → score sur 100

Une méthodologie

FMTI : Foundation Model Transparency Index - les sous-domaines

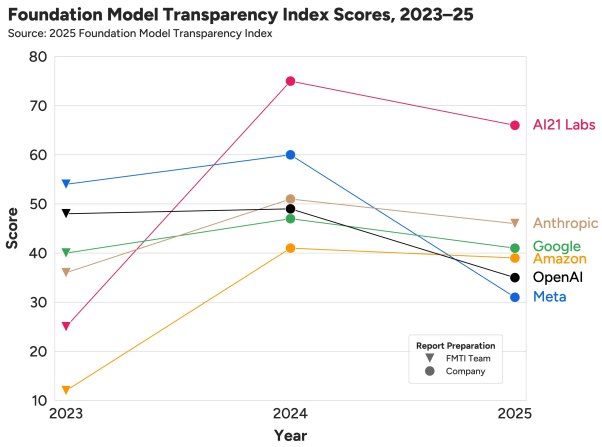


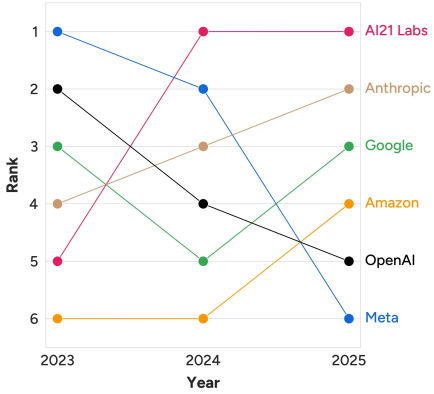
FMTI : Foundation Model Transparency Index - les résultats de l'étude de 2025



- Entreprises moins impliquées dans l'indice
- Toutes les entreprises ne sont pas devenues moins transparentes
- La baisse de la transparence n'est pas uniforme.

FMTI : Foundation Model Transparency Index - les résultats de l'étude de 2025





Les classements s'inversent entre 2023 et 2025

FMTI : Foundation Model Transparency Index - les résultats de l'étude de 2025

Foundation Model Transparency Index Scores by Major Dimensions of Transparency, 2025

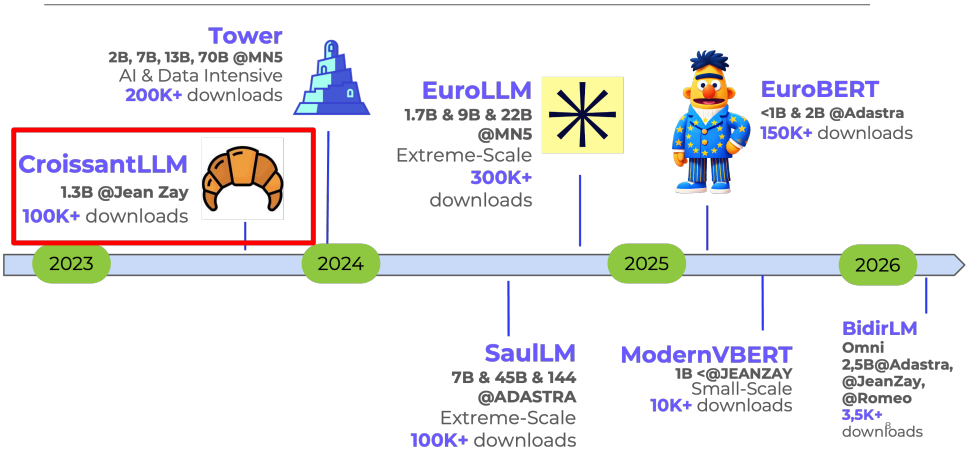
Source: 2025 Foundation Model Transparency Index

Major Dimensions of Transparency	AI21labs Jamba 1.6	Qwen 3	amazon Nova Premier	Claude 4	DeepSeek-R1	Google Gemini 2.5	IBM Granite 3.3	Meta Llama 4	Midjourney V7	Medium 3	OpenAI o3	WRITER Palmyra X5	Grok 3	Average
Data Acquisition	92%	17%	17%	25%	17%	33%	100%	33%	0%	0%	8%	58%	0%	31%
Data Properties	0%	20%	0%	0%	20%	0%	100%	20%	0%	0%	0%	40%	0%	15%
Compute	22%	11%	11%	0%	44%	11%	100%	22%	0%	0%	0%	100%	11%	26%
Model Information	75%	75%	0%	25%	75%	0%	100%	75%	0%	0%	0%	75%	0%	38%
Model Access	50%	50%	50%	50%	50%	50%	100%	50%	0%	25%	0%	50%	0%	40%
Capabilities	75%	50%	50%	25%	50%	25%	75%	50%	0%	25%	25%	50%	25%	40%
Risks	60%	0%	40%	60%	20%	20%	100%	20%	0%	0%	60%	40%	0%	32%
Model Mitigations	60%	0%	60%	80%	20%	40%	80%	0%	0%	20%	80%	40%	0%	37%
Release	88%	63%	75%	75%	63%	88%	100%	50%	63%	38%	63%	88%	50%	69%
Usage Data	20%	0%	20%	60%	0%	0%	80%	0%	20%	0%	20%	100%	0%	25%
Impact	71%	0%	0%	29%	0%	29%	86%	14%	29%	14%	14%	86%	0%	29%
Post-deployment Monitoring	71%	0%	57%	57%	0%	43%	100%	29%	0%	43%	71%	86%	0%	43%
Model Behavior Policy	100%	50%	75%	100%	75%	75%	100%	75%	25%	0%	75%	50%	75%	67%
Acceptable Use Policy	80%	60%	80%	100%	60%	80%	80%	40%	60%	60%	60%	80%	60%	69%
Downstream Mitigations	100%	40%	100%	100%	0%	100%	100%	80%	40%	80%	100%	100%	40%	75%
Average	**64%**	**29%**	**42%**	**52%**	**33%**	**40%**	**93%**	**37%**	**16%**	**20%**	**38%**	**69%**	**17%**	

Les données d'entraînement restent opaques - De nombreuses entreprises ne communiquent pas les informations de base concernant le modèle lui-même.

Petit retour d'expérience

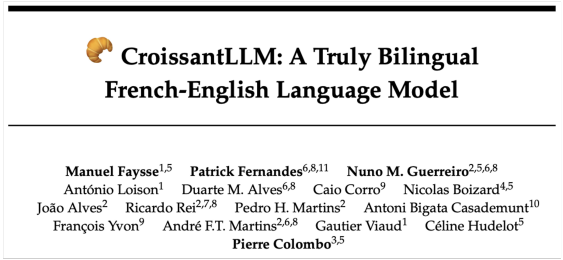
Une suite de modèles souverains



Petit retour d'expérience

CROISSANT LLM : un petit modèle de langue

Un partenariat industriel et universitaire pour un SLM souverain



CroissantLLM: A Truly Bilingual French-English Language Model

Manuel Faysse[1,5] Patrick Fernandes[6,8,11] Nuno M. Guerreiro[2,5,6,8]
António Loison[1] Duarte M. Alves[6,8] Caio Corro[9] Nicolas Boizard[4,5]
João Alves[2] Ricardo Rei[2,7,8] Pedro H. Martins[2] Antoni Bigata Casademunt[10]
François Yvon[9] André F.T. Martins[2,6,8] Gautier Viaud[1] Céline Hudelot[5]
Pierre Colombo[3,5]

CentraleSupélec ILLUIN TECHNOLOGY TÉCNICO LISBOA Unbabel cnrs Institut du Développement et des Ressources en Informatique Scientifique



Manuel Faysse
CentraleSupélec, MICS
PhD student



Pierre Colombo
CentraleSupélec, MICS
Assistant Professor



Céline Hudelot
CentraleSupélec, MICS
Professor



Nicolas Boizard
CentraleSupélec, MICS
Diabolocom
PhD Student

Petit retour d'expérience

Un projet de recherche avec des motivations industrielles



Recherche

CroissantLLM est un projet de recherche visant à étudier l'impact du **bilinguisme** sur le pré-entraînement et les performances des modèles de langue.



Industrie

Un modèle conçu pour être **suffisamment petit pour** fonctionner en local, mais suffisamment **performant** pour exécuter des tâches génératives souvent réservées à des modèles plus volumineux (**entraînement optimal pour l'inférence**). Il est entraîné uniquement sur des **données sous licence permissive.**



Open-Source

Basé sur l'open source, avec des modèles, des données, des bases de code et des critères d'évaluation **ouverts,** permettant ainsi aux **chercheurs et aux praticiens** d'en tirer profit.



Petit retour d'expérience

Les données

L'entraînement des LLMs nécessite d'énormes quantités de données... En français et sous licence permissive, il est encore plus difficile d'en rassembler suffisamment !

Corpus



Culture
Données sous licence permissive

- Livres dans le domaine public
- Podcasts
- Poèmes
- Paroles de chansons
- Sous titres de films



Business
Données de l'industrie et des administrations

- Corpus de lois
- Débats parlementaires
- Décisions administratives
- Documents commerciaux publics



Connaissance
Données scientifiques et factuelles

- Encyclopédies
- Manuels scolaires
- Résumés de thèses
- Publications scientifiques



Traduction
Données parallèles Anglais-Français

- Une quantité considérable de paires de traduction provenant de différents domaines
- Filtrées à l'aide de méthodes d'estimation de la qualité à l'état de l'art



Internet
Données filtrées du web

- Données à l'échelle du Web filtrées pour obtenir des textes français et anglais de haute qualité
- Code Github sous licence libre

Petit retour d'expérience

Les données

L'entraînement des LLMs nécessite d'énormes quantités de données... En français et sous licence permissive, il est encore plus difficile d'en rassembler suffisamment !

Corpus



Culture
Données sous licence permissive



Business
Données de l'industrie et des administrations



Connaissance
Données scientifiques et factuelles



Traduction
Données parallèles Anglais-Français



Internet
Données filtrées du web

Pré-traitement des données



Identification et scrapping → Déduplication → Filtrage → Suréchantillonnage



Petit retour d'expérience

Transparence & Open-Source

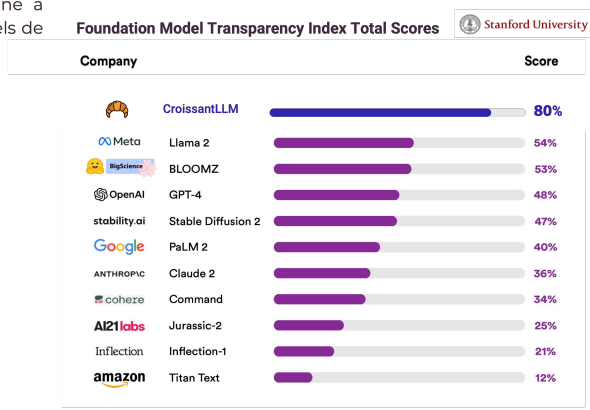
Projet fondé sur la transparence, destiné à servir de ressource utile aux professionnels de l'industrie et aux chercheurs !

Processus d'entraînement documenté du début à la fin

Ouvert à tous :

- **Corpus d'entraînement**
- **Points de contrôle du modèle**
- **Critères d'évaluation**
- **Bases de code**

Pas de restrictions d'usage (MIT)



Petit retour d'expérience

La transparence : un outil de recherche

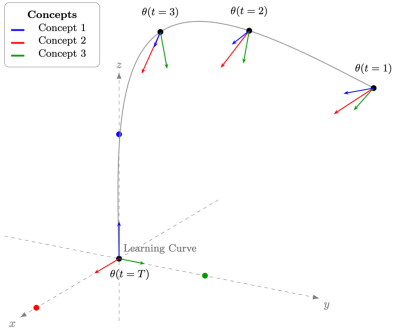


Figure 8: Evolution of θ over discrete time steps, with vectors indicating orientations toward three fixed concepts. The vector lengths represent the coefficients associated with each concept. The main intuition behind this figure is that, during training, the model may learn similar concepts but represent them differently (here, concept directions do not evolve linearly over time). In such a case, our process may fail to properly track concepts for direct comparison. However, the relative importance of concepts should remain stable-hence, we should focus on comparing their significance rather than their raw orientation (illustrated here by the arrow lengths).

Travail de Raphael Bernas avec Fanny Jourdan and A. Poché, IRT Saint Exupery



Interpreto

Interpretability Toolbox for LLMs

DEEL FOR ANITI



Message : concevoir des modèles transparents, souverains semble possible mais c'est surtout nécessaire

Conclusion - Ouverture

Ouverture — l'IA agentique change la nature du problème

	Modèle prédictif	**Agent**
Sortie	une prédiction	une **séquence d'actions**
Qui agit ?	l'humain, après lecture	le système, **directement**
Effets	réversibles (on ignore la sortie)	parfois **irréversibles** (paiement, envoi, suppression)
Périmètre	un modèle	modèle + outils + mémoire + autres agents
Ce qu'on explique	une décision	une **trajectoire**

Ce qui change

- L'humain passe de *décideur* à *superviseur*, voire à simple destinataire du résultat.
- Les erreurs se **composent** sur des dizaines d'étapes.
- Les entrées viennent du monde (pages web, fichiers, réponses d'outils) ⇒ **injection d'instructions**.

Enjeu

Plus l'autonomie est grande, plus la **transparence**, l'**explicabilité** et la **responsabilité** doivent être pensées *by design*.

Transparence et explicabilité : expliquer une trajectoire

Expliquer *pourquoi l'agent a agi ainsi*

- Les traces de raisonnement (ReAct, CoT) **ne sont pas des explications fidèles** : le modèle ne dit pas toujours ce qui a réellement guidé sa réponse.
- Il faut expliquer le **plan**, le **choix d'outil**, ce que l'agent **savait** à chaque étape, et pourquoi il s'est arrêté (ou non) pour demander.
- Interprétabilité mécaniste et concepts : piste pour vérifier ce qui est *utilisé*, pas seulement *verbalisé*.

Turpin et al., NeurIPS 2023 ; Chen et al., 2025 — *Reasoning models don't always say what they think*

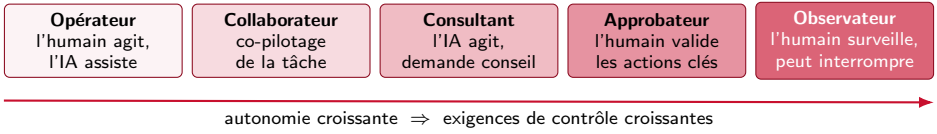
Rendre visible *ce que fait l'agent*

- **Journalisation** complète : prompts, appels d'outils, observations, sorties — l'inspection plutôt que la seule explication.
- **Identification** : savoir qu'on interagit avec un agent, et lequel (identifiant, déployeur, version).
- **Documentation** étendue : de la *model card* à la « fiche agent » (outils accessibles, permissions, limites, conditions d'arrêt).

Chan et al., FAccT 2024 — *Visibility into AI Agents* ; AI Act, art. 12 (registres) et art. 50 (transparence)

Une trace lisible n'est pas une explication fidèle : **pouvoir inspecter, pas seulement expliquer**.

Responsabilité : l'autonomie se règle, elle ne se subit pas



Niveaux d'après Feng, McDonald & Zhang, 2025 — *Levels of Autonomy for AI Agents*

Une chaîne de responsabilité qui s'allonge

- Fournisseur du modèle, concepteur de l'agent, fournisseurs d'outils, déployeur, utilisateur, *autres agents*...
- La responsabilité reste **humaine et organisationnelle** — l'agent n'est pas un agent moral.
- Risque de « *moral crumple zone* » : l'humain en bout de chaîne absorbe la faute sans avoir eu la maîtrise.

Elish, 2019 ; Dignum, 2019

Des leviers de conception

- Choisir le niveau d'autonomie **selon la tâche et le risque**, pas selon la capacité du modèle.
- Validation humaine avant les actions **irréversibles** ; moindre privilège ; budget ; arrêt d'urgence.
- Contrôle humain *effectif* (AI Act, art. 14) : l'humain doit comprendre assez pour pouvoir dire non.

Mitchell et al., 2025 — *Fully Autonomous AI Agents Should Not be Developed*

Messages à retenir

Ce qu'on a vu

- L'IA responsable s'appuie sur des **principes** (cadre ART) qu'il faut traduire en **outils**.
- L'explicabilité : du *où* (attribution) au *quoi* (concepts) — avec une exigence de **fidélité**.
- La transparence : propriété du **système et de son cycle de vie**, désormais une obligation légale (AI Act).

Avec les agents

- On explique des **trajectoires**, pas des prédictions isolées.
- La transparence passe par la **traçabilité** des actions et l'**identification** des agents.
- La responsabilité se conçoit en réglant l'**autonomie** et les points de **contrôle humain**.

Questions ouvertes

- Qui répond d'une action produite par l'interaction de *plusieurs* agents ?
- Comment garder un contrôle humain significatif quand les agents agissent plus vite et plus longtemps que nous ne pouvons suivre ?
- Peut-on certifier un système dont le comportement dépend des outils et des données qu'il rencontre en cours de route ?

Déléguer une tâche à un agent ne délègue pas la responsabilité.

Merci à toute l'équipe






















Many more